Estimating Sanitary Sewer Pipeline Infrastructure from Basic Characteristics of a Service Zone

by

Jessica May Winter

Thesis presented in partial fulfilment
of the requirements for the degree of
Master of Engineering
in the Faculty of Engineering at
Stellenbosch University.

Supervisors:

Mr Carlo Loubser

Mrs Adèle Bosman

March 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights, and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signed: …Jessica Winter…. Date: …March 2021……

Abstract

The standard detailed design and cost estimation for a sewer network involves considerable time and financial investment. There are, however, many cases where a rapid assessment of the sewer infrastructure or related costs associated with a service zone might be required. Accordingly, there have been numerous approaches to rapid sewer infrastructure assessment in the literature, ranging from the automated generation of entire sewer network plans to direct cost estimation methods. Yet, to date, no widely available tool has been developed that can be applied to reliably estimate the expected sewer pipeline infrastructure associated with a service zone in South Africa.

The main aim of this study was therefore to develop a method for estimating the sewer pipeline infrastructure required for a service zone, based on limited information, that could be applied to both existing and future developments.

In order to achieve the stated aim, a database of South African sewer network data was used in the development of three major study outcomes. Study Outcome I involved developing multiple linear regression models for estimating the total sewer pipeline length for a service zone using only basic service zone characteristics. Study Outcome II involved determining the average pipeline diameter distributions for different types of service zones, by which the total pipeline length could be disaggregated into lengths per diameter. Study Outcome III involved determining the average number of manholes per kilometre of sewer pipeline for different types of service zones, by which the total number of manholes for a service zone could be determined.

To satisfy Study Outcome I, models were developed for nine different categories of land use and area size using weighted least squares regression. The models allowed for estimation of the total pipeline length as a function of three variables, namely the service zone area size, relief (in terms of the difference between the mean elevation and the expected elevation of the network endpoint), and the density of contributing users (in terms of the number of unit hydrographs per hectare). The model strengths ranged from very good to moderate, with average percentage errors in the order of 10 – 35%. To satisfy Study Outcome II, pipeline diameter distributions were developed for 17 different categories of land use, area size and relief, which showed that the proportion of pipes with diameter ≤ 160 mm was always at least 90% for residential service zones, and at least 70% for non-residential service zones. To satisfy Study Outcome III, the average number of

manholes per kilometre of pipeline was determined for six different categories of land use and area size, which yielded an average manhole distribution in the order of 20 manholes/km.

Combined, the three study outcomes form an infrastructure estimation tool that enables reasonably reliable estimation of the sewer pipeline length per approximate diameter and the number of manholes associated with a service zone, applicable to service zones on a development scale smaller than 450 hectares.

Opsomming

Die gedetailleerde ontwerp en kosteberaming van rioolnetwerke is gewoonlik tydintensief en verg substansiële finansiële insette. Daar is egter baie gevalle waar 'n vinnige evaluering van die benodigde rioolpypleiding infrastruktuur, of die verwagte koste verbonde aan 'n dienstesone, benodig word. Vanuit die literatuur is daar talle benaderings vir die vinnige assessering van die benodigde rioolpypleiding infrastruktuur, wat wissel van die outomatiese generasie van volledige riooluitlegte tot direkte kosteberamingsmetodes. Desnieteenstaande, is daar nog geen algemeen beskikbare metode, waarvolgens die verwagte rioolpypleiding infrastruktuur vir 'n dienstesone in Suid-Afrika betroubaar beraam kan word nie.

Die hoofdoel van hierdie studie was dus om 'n metode te ontwikkel waarvolgens die rioolpypleiding infrastruktuur wat vir 'n dienstesone benodig word, beraam kan word, gebaseer op beperkte inligting. Die metode moes toepasbaar wees op bestaande sowel as toekomstige ontwikkelings.

Ten einde die studiedoelwitte te bereik, is 'n databasis van Suid-Afrikaanse rioolnetwerkdata gebruik vir die ontwikkeling van drie primêre uitkomste. Studie-uitkoms I behels die ontwikkeling van veelvuldige lineêre regressiemodelle vir die beraming van die totale rioolpyplengte vir 'n dienstesone, deur slegs basiese eienskappe van die dienstesone te gebruik. Studie-uitkoms II behels die bepaling van die gemiddelde verspreiding van die pyplyndiameter vir verskillende soorte dienstesones, waarvolgens die totale pypleidinglengte in totale lengte per pypdeursnee opgedeel kan word. Studie-uitkoms III behels die bepaling van die gemiddelde aantal mangate per kilometer rioolpypleiding vir verskillende soorte dienstesones, waarvolgens die totale aantal mangate vir 'n dienstesone bepaal kan word.

Vir nege verskillende kategorieë van area en grondgebruik, is regressiemodelle ontwikkel waarvolgens die totale pypleidinglengte bereken kon word. Daar is bevind dat die totale pypleidinglengte 'n funksie is van drie veranderlikes, naamlik die grootte van die dienstesone, reliëf (in terme van die verskil tussen die gemiddelde hoogte en verwagte hoogte van die riool eindpunt), en die digtheid van gebruikers (in terme van die aantal hidrografiese eenhede per hektaar). Die akkuraatheid van die modelle het gewissel van baie goed tot redelik, met 'n gemiddelde persentasie fout tussen 10 en 35%. Vir 17 verskillende kategorieë van grondgebruik, oppervlakte en reliëf, is pyplyn-deursnee-verdelings ontwikkel. Daar is bevind dat vir nie-residensiële dienstesones, meer as 70% van alle pype ≤ 160 mm in deursnee is. Vir residensiële

dienstesones daarenteen, het meer as 90% van alle rioolpype 'n deursnee van ≤ 160 mm. Laastens, vir ses verskillende kategorieë van grondgebruik en oppervlakte, is bevind dat die gemiddelde getal mangate per kilometer pypleiding in die orde van 20 mangate/km is.

Deur gesamentlike toepassing van al drie studie-uitkomste, is 'n metode ontwikkel waarvolgens rioolpypleiding infrastruktuur tot 'n redelike mate van akkuraatheid beraam kan word, vir dienstesones kleiner as 450 hektaar.

Acknowledgements

Mr Carlo Loubser – thank you for entrusting me with this topic which was your own original idea; for your valuable technical guidance; and mostly for being a wonderful supervisor whose enthusiasm and encouragement gave me confidence in my own ability.

Mrs Adèle Bosman – thank you for taking me on and for investing your time and expertise in my work; your guidance gave me clarity in the most challenging parts of this study and your encouragement and kindness were very much appreciated.

To the people at GLS Consulting and especially Mark Hoppe, Johann Rudolph, Jurie Van Der Merwe and Erik Loubser – thank you for providing the data without which this study would not have been possible; and for welcoming me into your office and sharing your time, interest and invaluable expertise with me.

And to Frans Grotepass – thank you for laying the foundation for this study with the work you did for potable water networks.

Table of Contents

PART 1 – MAIN REPORT

List of Figures

List of Tables

List of Abbreviations, Symbols and Acronyms

AADD	Annual average daily demand (kL/d)
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
DHS	Department of Human Settlements (South Africa)
DEM	Digital elevation model
DWS	Department of Water and Sanitation (South Africa)
EE	Equivalent erven
GIS	Geographic Information System
GLS	GLS Consulting (Pty) Ltd
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MASL	Metres above sea level
OLS	Ordinary least squares
PCA	Principal component analysis
PDDWF	Peak daily dry weather flow (kL/d)
RMSE	Root mean squared error
UH	Unit hydrograph
WLS	Weighted least squares
WWTP	Wastewater treatment plant

A	Area of service zone polygon
H_{max}	Highest point elevation of the service zone
H_{mean}	Mean elevation of the service zone
H_{min}	Lowest point elevation of the service zone
L	Service zone length
P	Perimeter of service zone polygon
X_c	Centroid x-coordinate of bounding rectangle of service zone
X_{mouth}	X-coordinate of end manhole of the sample network
Y_c	Centroid y-coordinate of bounding rectangle of service zone
Y_{mouth}	Y-coordinate of end manhole of the sample network
Z_{mouth}	Z-coordinate of end manhole of the sample network

PART 1 – MAIN REPORT

Chapter 1
INTRODUCTION

1.1 Background

Since the time of the Ancient Greeks, underground channels or sewers have been used to convey human waste away from inhabited areas. For centuries, sewer systems typically discharged wastewater into water bodies for dilution, onto crop fields for irrigation and fertilisation, or onto unused land. It was only in the late 19th century that the first cities began to treat wastewater prior to discharge using sedimentation and chemicals (Lofrano & Brown, 2010). In South Africa, the first flush toilet connected to a wastewater conveyance system was installed in 1884. The country's major cities – Cape Town, Johannesburg, Bloemfontein, and Durban – only gained access to waterborne sanitation via sewers at the start of the 20th century (van Vuuren & van Dijk, 2011; Mäki, 2007).

Modern sanitary sewer systems consist of a network of pipes which convey wastewater by means of gravity and pumping to a centralised wastewater treatment facility, where it is treated to an acceptable standard before being discharged to a suitable location. Currently, only around 80% of South African households have access to improved[1] sanitation (Statistics South Africa, 2018), although it is intended that connection to the public sewerage network should ultimately be provided as the national basic level of service for sanitation (SAICE, 2017).

To determine the sewer network infrastructure required for a particular service zone, a detailed hydraulic design process is required. Typically, the erf layout, road layout and topography are used to establish the sewer network layout. Thereafter, hydraulic input parameters such as the expected flows and ground slopes along pipeline sections are used in the hydraulic design to determine the pipe sizing, placement of manholes and special structures, and the necessity of

[1] Improved sanitation is defined as access to at least a ventilated improved pit toilet, or to a flush toilet connected to a public sewerage network or septic tank (Statistics South Africa, 2018)

pumps and rising mains. When the hydraulic design is complete, other project information such as the required excavation and pipeline bedding can be determined.

1.2 Problem Statement

The standard detailed design and cost estimation for a sewer network involves considerable time and financial investment. There are many cases where a rapid assessment of the sewer infrastructure or associated costs might be required, such as in a feasibility study for a proposed development, or for infrastructure management and cost projection on a town planning level. Therefore, the ability to quantify sewer pipeline infrastructure associated with a service zone based on limited information holds considerable value for project planning.

Accordingly, there have been many approaches in the literature to estimating sewer pipeline infrastructure, from the automated generation of entire sewer networks, to direct costing methods. There have also been many methods devised globally for quantifying the expected components of a sewer network, differing in their approaches and intended applications. However, to date, no widely available tool has been developed that can be applied to reliably estimate the expected sewer pipeline infrastructure associated with a service zone in South Africa.

1.3 Research Objectives

The main aim of this study is to develop a method for estimating the sewer pipeline infrastructure required for a service zone, based on limited information, that can be applied to both existing and future developments. This aim necessitated three major study outcomes, namely:

- Study Outcome I: The development of a model for estimating the total sewer pipeline length for a service zone using basic service zone characteristics.
- Study Outcome II: The development of pipeline diameter distributions for disaggregating the total pipeline length into lengths per diameter, for different types of service zones.
- Study Outcome III: The quantification of the typical number of manholes required along a length of pipeline, for different types of service zones.

The chosen approach was to develop the three study outcomes listed above statistically using South African sewer network data.

The following objectives were set in order to meet the main aim of the study:

- To identify the potentially-influential characteristics that would reasonably be known for existing or planning-stage future developments.
- To source a suitable South African sewer network database, and to extract from this a set of sample networks for which the potentially-influential characteristics are quantified.
- To use appropriate statistical methods to develop the required study outcomes.
- To evaluate the accuracy and reliability of the tools arising from the study outcomes, and thus their suitability for application.

1.4 Motivation for the Study

While there are existing tools for feasibility-stage costing of sewer projects, many of them require an assumption to be made regarding the expected pipeline infrastructure, particularly in terms of the total pipeline length per diameter or material. A tool that enables the expected pipeline infrastructure to be reliably estimated could therefore offer considerable benefits for improving the accuracy of the cost estimations that can be made using existing costing methods.

Furthermore, the envisioned pipeline infrastructure estimation tool would also have potential in non-costing applications, such as:

- Updating infrastructure databases where information is outdated, lost or confidential.
- Serving as a design benchmark for new sewer schemes.
- Aiding in preliminary wastewater treatment plant (WWTP) sizing calculations by allowing for more accurate infiltration estimates.
- Providing more detailed information for decision-making when comparing a traditional WWTP and sewer network to more modern decentralised solutions.
- Helping urban planners to determine the wastewater network size that achieves optimal economies of scale.

1.5 Assumptions

The three major study outcomes are developed using analyses of existing infrastructure, therefore the inherent assumption on which the study rests is that the existing networks have been designed optimally and function adequately. To manage this assumption, appropriate steps have been

taken, particularly in the data collection process where the data quality and the sample selection are concerned. With such steps in place, this can be considered a reasonable assumption.

1.6 Delineations and Limitations

This study aims to determine the expected sewer pipeline infrastructure associated with a service zone only in terms of the pipeline length per diameter and the number of manholes. The occurrence of any other structures is not considered in the study. This includes the presence of rising mains, which is dependent on the specific layout of a sewer network and was therefore considered too specific a factor to predict statistically. The infrastructure estimation tool developed in this study is therefore only applicable in gravity-driven catchments.

The infrastructure estimation tool developed in this study is intended for application in existing and new developments. Therefore, the sample networks used to develop the tool represented networks on a development and suburb scale. Consequently, the results account mainly for reticulation and collector sewers rather than bulk lines, and the tool is not suitable for application on a town or city scale. Furthermore, no allowance is made for outside flow contributions from adjacent upstream developments draining through the development of interest, and the tool is only applicable to developments on the upstream end of a catchment.

1.7 Definitions

In the context of this study:

- A ‘service zone’ is defined as the development or suburb-scale region serviced by the sewer network, with an associated number of connected users per land use.
- A ‘sample network’ is defined as a system of gravity sewer pipes converging to a single endpoint at the first manhole that receives the full combined flow from the service zone.
- ‘Limited information’ refers to information that can reasonably be assumed to be known at the planning or feasibility stage of a proposed sewer project. This includes the service zone boundary, number of equivalent erven per land use, the location of the sewer end point, and the topography in the form of a digital elevation model (DEM) or similar.
- The ‘infrastructure estimation tool’ refers to the method for estimating the required sewer pipeline infrastructure for a service zone that is developed in this study.

1.8 Thesis Structure

Following the introductory chapter, this thesis comprises nine chapters. The order and purpose of each chapter is summarised in Figure 1-1. The thesis begins with the Literature Review in Chapter 2, followed by the Research Design in Chapter 3. The following five chapters detail the methodology, namely the Data Collection in Chapter 4, the Regression Methods in Chapter 5, and the Analysis for Study Outcome I, Analysis for Study Outcome II and Analysis for Study Outcome III in Chapter 6, Chapter 7 and Chapter 8, respectively. The Results are presented and discussed in Chapter 9, followed by the Conclusion in Chapter 10.

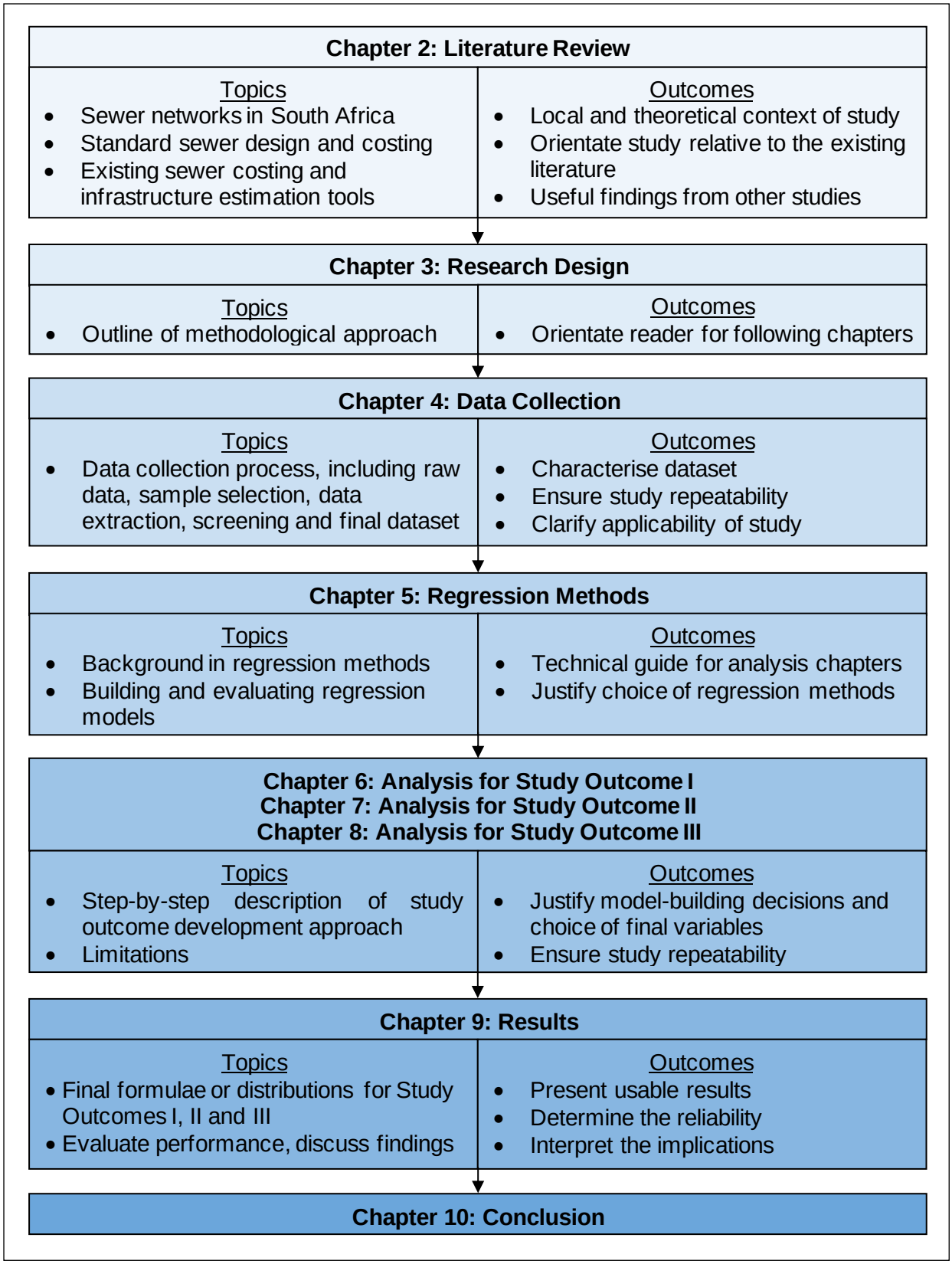


Figure 1-1: Thesis structure.

Chapter 2
LITERATURE REVIEW

The Literature Review has three focus areas. Firstly, to provide the context surrounding sewer networks in South Africa. Secondly, to explain the broad theory basis of standard sewer design and costing practices in South Africa. Thirdly, and most importantly, to review the various methods in the literature for estimating the expected layout, cost, or infrastructure components of a sewer network using limited information. Aside from providing the local and theoretical context of the study, the two major intended outcomes of this chapter are:

- To orientate the proposed study relative to the surrounding literature, by highlighting the gap that it aims to address.
- To highlight the findings from other studies that influenced the methodology of the proposed study.

This chapter is structured in five main sections. The first provides a summary of the history and current state of wastewater networks in South Africa. The second provides an overview of the standard design and costing of sewer networks, according to South African guidelines. The final three sections provide discussions of the major subsets of the literature related to early-stage sewer infrastructure estimation, namely automated generation of sewer network plans, direct cost estimation of sewer networks, and estimating the infrastructure components of sewer networks.

2.1 Summary of Wastewater Networks in South Africa

This section provides the local context of wastewater networks through a discussion of the history, current extent, future expansion, and conveyance system type of South African sewer networks.

2.1.1 History

While there is evidence of people using drainage channels for houses and public latrines as early as 3500 BC, the development of more advanced sewer practices is normally accredited to the

ancient Greeks. Public latrines drained via pipes to a single collector sewer, which conveyed waste- and stormwater to nearby agricultural fields for irrigation (Lofrano & Brown, 2010). By comparison, the history of the South African sewerage network is less than 150 years long. The first flush toilet connected to a waterborne sanitation system in South Africa was installed in the Karoo town of Matjiesfontein in 1884 (van Vuuren & van Dijk, 2011).

The development of sewerage infrastructure in South Africa was largely dependent on the water supply, as is illustrated in the case of Cape Town, the country's oldest city. As described by Juuti et al. (2007), at the start of the 18th century, the city obtained its limited water via channels from mountain streams and springs. People collected their water in buckets or from one of the few public water fountains. In the early 18th century, water supply pipelines were constructed, mostly out of wood or lead. In the early 19th century, cast iron pipes were constructed to supply three major streets after the completion of a reservoir in 1811. With the growing population and limited storage, water was still in short supply, and a more extensive water supply network only began construction in the 1860's.

In the absence of a reliable water supply for flushing, sanitation lagged far behind (Mäki, 2007). According to Juuti et al. (2007), the 18th- and early 19th-century sanitation practices in Cape Town consisted mainly of the emptying of sanitary buckets onto wasteland, into the bay, or simply into the streets. In the 1850's, the main sewer of the city was a wide, uncovered drain running into the bay (Mäki, 2007). With the completion of the Molteno water reservoir in 1886, full waterborne sanitation could be considered as a feasible option. The sewer network plans were, however, only finalised in 1896, and the work took another 10 years to complete. Finally, in the early 1900's, Cape Town had a functional sewer system with a marine outfall at Green Point (Mäki, 2010).

According to Mäki (2010), the other major cities began to improve their sanitation at around the same time. The Durban sewer system became operational in 1896. Johannesburg built a sewerage system for the town centre in 1903, and Port Elizabeth also began its first large drainage scheme that same year. Bloemfontein only began to improve its sewerage conditions when prompted by the Spanish influenza pandemic in 1918; but by 1924 all of 'white' Bloemfontein was connected to the waterborne sewerage system (at the time, racial segregation policies were in place).

Towards the end of that century, when South Africa transitioned out of the Apartheid era in 1994, an estimated 20.5 million people (51% of households) still lacked access to basic sanitation (WWF-SA, 2016). The National Water Services Act of 1997 addressed this, stipulating that "everyone has a right of access to basic water supply and basic sanitation". Basic sanitation was defined as "the prescribed minimum standard of services necessary for the safe, hygienic and adequate collection, removal, disposal or purification of human excreta, domestic wastewater and sewage from households, including informal households" (South Africa, 1997). By 2002, the percentage of households with access to improved sanitation stood at 61.7%, where improved sanitation refers to access to a flush toilet connected to a public sewerage system or septic tank, or at least a ventilated improved pit (VIP) toilet (Statistics South Africa, 2018).

2.1.2 Current extent

Moving to the present day, there are 152 Water Services Authorities in South Africa providing wastewater services via a network of 824 wastewater collector and treatment systems (WWF-SA, 2016). The most recent South African census data from 2018 indicate that, nationally, 83% of households now have access to improved sanitation (Statistics South Africa, 2018). This access is, however, unevenly distributed, with less than 50% of households having access to improved sanitation in the provinces of the Eastern Cape, KwaZulu-Natal, North West, Limpopo, and Mpumalanga. Furthermore, as suggested by Statistics South Africa (2018), access to improved sanitation appears to have stabilised around 80% over the past few years.

2.1.3 Future expansion

According to SAICE's 2017 Infrastructure Report Card for South Africa (SAICE, 2017), political pressure to provide connection to the public sewerage network as a basic level of sanitation is heavily impacting the cost of service provision, thereby slowing service delivery down. Another obstacle is that the state of wastewater treatment plant (WWTP) infrastructure, which is generally satisfactory in major urban areas, is a matter of "grave concern" outside of these centra (SAICE, 2017). Up to 30% of all WWTPs are in critical condition, resulting in increasing quantities of poorly treated or untreated wastewater being discharged into streams. Thus, for the sewerage infrastructure network to continue to expand in a sustainable manner, considerable investment will be required in both pipeline and WWTP infrastructure.

However, waterborne sanitation is not in all instances a feasible solution. Universal access to waterborne sanitation is beyond the financial reach of most developing countries, particularly in sparse rural settlements (WWF-SA, 2016). Alternative sanitation solutions that are implemented in South Africa include various on-site systems such as the ventilated improved pit (VIP) toilet, composting toilet, urine-diverting dry toilet, low-flush toilet, pour-flush toilet, aqua-privy, septic leach field system, and anaerobic digestor; as well as basic sewer types such as vacuum, small-bore and simplified (or "shallow") sewer systems (DHS, 2019).

2.1.4 Separate sanitary sewer and stormwater system

Worldwide there are two major types of sewer systems, namely combined and separate sewers. A combined sewer system conveys both wastewater and stormwater runoff to a WWTP. When designing such systems, stormwater produces significantly higher flow peaks than wastewater, and normally dominates the design. In a separate sewer system, wastewater is conveyed via one converging network to a WWTP; and stormwater via a different network to a safe discharge site such as a river. The separate wastewater network is designed to prevent stormwater ingress, although some still enters through places such as uncovered manholes and illegal connections. The vast majority South Africa's sewerage network consists of separate systems (Stephenson & Barta, 2005), and all new systems are designed this way.

2.1.5 Concluding summary

Overall, the South African sewer system is relatively young and requires significant development and expansion. In addition to urban growth, much of the network expansion will involve upgrading the sanitation level of service in low income areas, and connecting rural service zones to new waterborne sanitation, although this progress will be slow. The next section provides an overview of the standard methods used when designing new sewer networks in South Africa.

2.2 Standard Design and Costing of Sewer Networks in South Africa

The purpose of this section is to ensure that the standard sewer design and costing methods are firmly understood before methods for circumventing this process can be considered; as well as to provide an adequate theoretical background for the methodological decisions that were made in this study. This section is largely a summary of the standard South African guideline for the design of wastewater systems, namely 'Section K: Sanitation' of 'The Neighbourhood Planning and Design Guide'. This guideline was produced by the Department of Human Settlements or DHS (2019), and is commonly known as the 'New Red Book'. Therefore, this section will be of more value to the general reader than to the advanced reader, for whom this information might be common knowledge.

In the following subsections, the topics covered include sewer network layout, calculation of design flow, pipe sizing and network design criteria, an example of a typical design software program, and cost estimation. The design of rising mains and pumps is omitted as it is beyond the study scope.

2.2.1 Sewer layout

Before detailed hydraulic design of a sewer network can take place, a plan of the proposed sewer layout and the erven serviced at each connection is required. This planning is often performed in conjunction with the planning of other engineering services and the overall service zone layout. Usually, municipalities stipulate their own layout guidelines and specifications, but there are some general considerations that should be followed (DHS, 2019). Firstly, the sewer layout should be chosen to ensure the most economical design, considering the topography. Pipelines should follow the natural gradient of the ground, while avoiding gradients that are too flat or steep. Sewer pipes should be in locations where they are easily accessible, such as open areas, road reserves, or municipal land; and should be laid next to properties where they give the most direct benefit. Road crossings should be avoided where possible, and there should be minimum interference with existing structures and services.

2.2.2 Design flow

In order to determine the flow originating from the various erven serviced by the connections to the network, certain information is required. Firstly, the water usage hydrographs for the different land uses serviced are required. The unit hydrographs (UHs) provide the 24-hour water usage pattern for each land use, and it is assumed that the wastewater production follows the same pattern. UHs have a peak flow of one, and corresponding percentage waste and peak factors by which the UH values can be multiplied to obtain the actual expected hourly wastewater production. Estimates of the leakage and base flows are provided as accompanying constants. Ideally, municipality-specific hydrographs should be used, but more generally-applicable ones are also available. If available, the known water usage in terms of the annual average daily demand (AADD) should be obtained, as well as the estimated percentage of the AADD that contributes to the wastewater flow. Furthermore, municipality-specific estimates for the infiltration rate, stormwater ingress allowance and peak day factor should be obtained where possible.

The design flow for sewers is the flow rate that the system should be able to convey without surcharging. It is made up of three components, namely regular flow, infiltration, and stormwater ingress. Regular flow or local inflow consists of the sewage return flow from domestic and commercial water users, as well as base flow and leakages from the plumbing system. Infiltration is the groundwater that seeps into the pipes through pipe joints and cracks. Stormwater ingress is the rainfall runoff that enters the system through uncovered manholes and illegal connections. In the subsections to follow, regular flow, infiltration and stormwater ingress are explored in further detail, including how these are used to calculate the following flows of interest:

- Peak Daily Dry Weather Flow or PDDWF (total flow for the peak day in a week, in kL/d)
- Average Daily Dry Weather Flow or ADDWF (total average daily flow, in kL/d)
- Instantaneous Peak Dry Weather Flow or IPDWF (L/s)
- Instantaneous Peak Wet Weather Flow or IPWWF (L/s)

2.2.2.1 Regular flow

Depending on the information available, there are three methods for calculation of the regular flow, as specified by DHS (2019). These are the unit hydrograph (UH) method, the AADD method, and the peak factor method. The peak factor method is the most approximate in nature, therefore only the more accurate UH and AADD methods are discussed here. The UH method is used if

the AADD is not known. Using the UH method, for a given number of equivalent erven (EE) per land use at a connection, the local inflow hydrograph at that connection is determined by multiplying the UH ordinates by the peak flow and adding the constant leakage, then multiplying by the number of equivalent erven. This is summed for the different land uses. If the AADD and percentage AADD wastewater contribution are known, then the AADD method can be used. It is similar to the UH method, but the UH method's output is scaled so that the volume of the inflow hydrograph is equal to the AADD sewer inflow, rendering it theoretically more accurate (GLS Consulting, 2019). These two methods are used to determine the regular flow rate at a given time as follows (DHS, 2019):

For one unit of a certain land use, Equation 2-1 defines the sewer inflow at time t. Equation 2-2 defines the total flow volume originating from that unit in a day.

$$HQ_t\left(\frac{\frac{L}{min}}{unit}\right) = UH_t \times Peak + Leak \qquad 2\text{-}1$$

$$Unit\ PDDWF\left(\frac{\frac{kL}{d}}{unit}\right) = \frac{1}{24}\sum_{t=1}^{24} HQ_t \times (24h \times 60\ min) \div 1000\ \mathrm{L} \qquad 2\text{-}2$$

Where:

HQ_t	Unit flow at a specific time step for land use type (L/min/unit)
UH_t	Unit hydrograph value for land use type at time t
$Peak$	Hydrograph peak flow for land use type (L/min/unit)
$Leak$	Hydrograph leakage flow for land use type (L/min/unit)

For the UH method, Equation 2-3 defines the flow rate originating from all units of a certain land use at time t. For the AADD method, Equation 2-4 defines the flow rate originating from all units of a certain land use at time t.

$$TQ_t\left(\frac{L}{s}\right) = HQ_t \times \left(\frac{Unit\ PDDWF\ \left(\frac{\frac{kL}{d}}{unit}\right) \times 1000L\ \div (24\ h\ \times 60\ min)}{\frac{1}{24}\sum_{t=1}^{24} HQ_t} \right) \times EE \div 60sec \quad \text{2-3}$$

$$TQ_t\left(\frac{L}{s}\right) = HQ_t \times \left(\frac{AADD\ \left(\frac{\frac{kL}{d}}{unit}\right) \times \frac{Ratio}{100} \times 1000L \div (24\ h\ \times 60\ \text{min})}{\frac{1}{24}\sum_{t=1}^{24} HQ_t\ \left(\frac{\frac{L}{min}}{unit}\right)} \right) \times EE \div 60sec \quad \text{2-4}$$

Where:

TQ_t	Calculated flow at a specific time step from all units of a land use type (L/s)
EE	Number of units of the land use type
$AADD$	Unit AADD for land use type (kL/d/unit)
$Ratio$	Portion of AADD that enters the sewer (%)

The regular flow can then be expressed as instantaneous peak dry weather flow, peak daily dry weather flow, or average daily dry weather flow using Equations 2-5, 2-6, and 2-7, respectively. Infiltration and ingress have not yet been accounted for.

$$IPDWF\ (excl.infiltration)\left(\frac{L}{s}\right) = \max\ (TQ_{t1,24}) \quad \text{2-5}$$

$$PDDWF\ (excl.infiltration)\left(\frac{kL}{d}\right) = EE\ \times Unit\ PDDWF \quad \text{2-6}$$

$$ADDWF\ (excl.infiltration)\left(\frac{kL}{d}\right) = \frac{PDDWF(excl.infiltration)}{Peak\ day\ factor} \quad \text{2-7}$$

2.2.2.2 Groundwater infiltration

Since groundwater enters sewers through cracks and joints in the pipes, the infiltration rate is dependent on the length and outside diameter of the pipe. In South Africa, a constant groundwater infiltration rate of 0.03 to 0.04 L/min/m pipe/m Ø is normally allowed for (DHS, 2019). It is noted that verification of the infiltration rate falls outside the scope of this study. For a specific point in the network, the infiltration originating from upstream pipes should be converted to the correct units and added to the IPDWF, PDDWF and ADDWF, so that it is now included in these flows.

2.2.2.3 Stormwater ingress

Rather than quantifying the volume of stormwater ingress, pipes are designed to have a certain percentage spare capacity to allow for stormwater ingress. The required spare capacity differs between municipalities, but DHS (2019) recommends 30% for reticulation, and 15 – 30% for outfall sewers. The spare capacity may be specified as absolute or relative. The absolute spare capacity is defined in Equation 2-8, and the relative spare capacity in Equation 2-9.

$$Absolute\ spare\ capacity\ (\%) = \frac{(Full\ flow\ capacity - IPDWF)}{Full\ flow\ capacity} \times 100 \qquad 2\text{-}8$$

$$Relative\ spare\ capacity\ (\%) = \frac{(Full\ flow\ capacity - IPDWF)}{(Full\ flow\ capacity - Upstream\ pump\ flow)} \times 100 \qquad 2\text{-}9$$

2.2.2.4 Determining peak flow

At any point in the network, after including the upstream regular flow and infiltration in the IPDWF at that point, the instantaneous peak wet weather flow or design flow is calculated by accounting for the spare capacity, as depicted in Equation 2-10.

$$IPWWF\ or\ Design\ Flow\ \left(\frac{L}{s}\right) = \frac{IPDWF\ (\frac{L}{s})}{(1 - Required\ spare\ capacity)} \qquad 2\text{-}10$$

2.2.2.5 Flow routing through the network

With modern software, whole unit hydrographs can be routed through the network so that the flow at any point in the network, at any time, can be viewed. The combined hydrograph peak at any point in the system represents the IPDWF at that point, and for pipe sizing this peak can simply be augmented by the required spare capacity. According to Yen, et al. (1976), there are several methods for modelling the flow through a sewer network, varying in complexity and accuracy.

The most basic methods simply sum the peak design flow or inflow hydrographs, ignoring time lag effects as well as the unsteady and non-uniform nature of the flow. Lag time can be accounted for by using the Manning or Darcy-Weisbach equations to determine the flow velocity, assuming full-flow conditions. From the flow velocity, the lag time is calculated and used to stagger the inflow hydrographs before summing them. The Chicago Hydrograph method, Illinois Urban Drainage Area Simulator (ILLUDAS), and the Transport and Road Research Laboratory (TRRL) method are all based on this principle (Yen, et al., 1976).

More complex methods simulate non-uniform flow and partial backwater effects, such as the nonlinear kinematic wave method and the Environmental Protection Agency Storm Water Management Tool (EPA SWMM). The most comprehensive method, the Illinois Storm Sewer System Simulation Model (ISS), uses dynamic flow equations to simulate unsteady and non-uniform flow, including both upstream and downstream backwater effects, as well as junction and manhole effects. However, the dynamic flow equations approach is computationally intensive, and the time-shifting or nonlinear kinematic wave approaches are sufficiently accurate for design purposes (Yen & Sevuk, 1975).

2.2.2.6 Determining flow velocity

The velocity associated with a flow rate is dependent on the chosen diameter, slope and material. It may be calculated using any of the equations presented in Table 2-1, provided that the results are close to those produced by the Colebrook-White formula (DHS, 2019). These equations assume uniform and unpressurised flow. To calculate the velocity, full pipe flow is assumed, and the selected equation from Table 2-1 is used to calculate Q_0 (full pipe discharge) and V_0 (full flow velocity). Then, with the design flow representing Q, the partial flow diagram in Figure 2-1 is used to obtain the flow velocity V.

Table 2-1: Equations to determine uniform flow velocity in sewer pipes (DHS, 2019).

Method	Formula	Roughness Coefficient
Chèzy	$Q = VA = 18\log\left(\frac{12\ R}{k_s}\right)A\sqrt{RS_o}$	k_s = 0.600
Colebrook-White	$\frac{1}{\sqrt{f}} = -2\log_{10}(\frac{k_s}{3.7\ D_H} + \frac{2.51}{Re\sqrt{f}})$	
Manning	$Q = VA = \frac{1}{n}AR^{\frac{2}{3}}S_0^{\frac{1}{2}}$	n = 0.012 Dependent on pipe material and condition.
Kutter	$C = \frac{23 + \frac{0.00155}{S_0} + \frac{1}{n}}{1 + \frac{n}{\sqrt{R}(23 + \frac{0.00155}{S_0})}} \times \sqrt{RS}$	

Where:

A Cross-sectional flow area (m^2)

R Hydraulic radius, or area divided by wetted perimeter (m)

k_s Absolute roughness of pipe interior (m)

S_0 Slope (m/m)

f Darcy-Weisbach friction factor

D_H Hydraulic diameter, or inside diameter for a circular pipe (m)

Re Reynolds' Number

n Manning's roughness coefficient

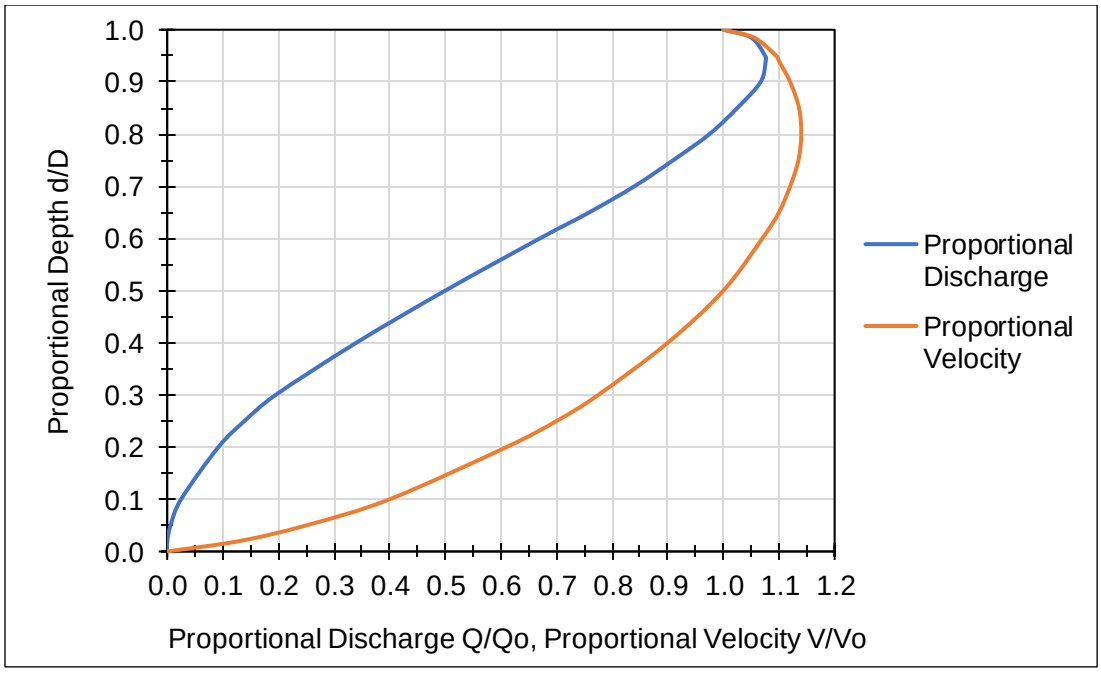


Figure 2-1: Partial flow diagram (DHS, 2019).

2.2.3 Design criteria

In addition to ensuring that the full flow capacity of each pipe is greater than its design flow, there are several other criteria that govern the design of sewer pipelines and networks. The most important ones from DHS (2019) that affect pipe diameter or gradient briefly discussed here.

2.2.3.1 Minimum and maximum flow velocity

To ensure sediments are flushed regularly, a minimum full-flow velocity of 0.6 – 0.7 m/s must be maintained in all gravity mains. For each pipe diameter, there is a corresponding minimum gradient to ensure the minimum full-flow velocity is met. Sewer pipes at the upper ends of sewer networks service fewer properties and flow full less regularly, and therefore should be steeper than the minimum gradient to ensure pipes are regularly cleared.

To prevent damage to pipelines, a maximum full-flow velocity of 2.5 m/s should not be exceeded. Higher velocities of up to 4 m/s may be acceptable for short periods in short pipe sections, as permitted by the pipe specifications (DHS, 2019).

2.2.3.2 Minimum diameters

To prevent blockages, minimum diameters are also specified. These may vary for different municipalities, but DHS (2019) recommends minimum diameters of 100 mm for reticulation, 150 mm for municipal sewers, and 200 mm in CBD zones to allow for future densification.

2.2.3.3 Pipe materials

Pipelines in South African sewer networks are made of various materials, including uPVC (un-plasticised polyvinyl chloride), HDPE (high density polyethylene), vitrified clay, reinforced concrete, fibre cement, brick, cast-iron, and steel. According to DHS (2019), it is now generally accepted that heavy-duty uPVC pipes should be used for diameters less than 400 mm, and reinforced concrete pipes should be used for diameters greater than 400 mm. Furthermore, only HDPE pipes should be used in areas underlain by dolomite. However, other materials that may also be considered are vitrified clay, fibre cement, cast-iron and steel.

2.2.3.4 Manhole spacing

Manholes should be placed at all significant points such as main sewer junctions, gradient changes, direction changes, diameter changes, junctions of more than two pipes, and near road crossings. Again, municipalities usually have their own specifications, but DHS (2019) recommends that the distance between manholes should not exceed 100 – 150 m where power-rodding equipment is available. Where only hand-operated rodding equipment is available, then it should not exceed 100 m. This maximum spacing should be decreased for steep gradients to ensure that the pressure head at any point does not exceed 6 m during a blockage. The maximum spacing may be increased along larger-diameter collector or outfall sections.

2.2.4 Modelling software example: Sewsan 6

There are many software packages available for the modelling of sewer networks. Sewsan is a computer application for the simulation and analysis of flow in sanitary sewer systems, developed locally by GLS Software (PTY) Ltd. In Sewsan, the sewer network model is embedded within a GIS database and displayed on a map. Aerial photographs can be loaded in the background to facilitate data capturing and analysis.

2.2.4.1 Functionality

Sewsan has three essential functionality modules (GLS Software, 2020). The Capturing module allows models to be built. The Analysis module simulates flow through the network and returns results such as flow volumes, peaks, velocities, and pressures throughout the system. This allows the user to verify compliance with hydraulic design criteria, or to identify bottlenecks and overflows. The Planning module can be used to run a so-called planning analysis, in which the diameters of pipes of insufficient capacity are iteratively increased until there are no more bottlenecks, and the user-defined spare capacity is accommodated in all pipes.

2.2.4.2 Flow generation

To model the network flows, the user uploads the required unit hydrograph information for each land use to the model. At each network node, the number of land parcels or unit hydrographs per land use are assigned, and the AADD specified. To calculate the regular flow, Sewsan follows the calculation method described in Section 2.2.2, and supports both the AADD and unit hydrograph methods. The infiltration is specified as a constant rate per pipe. Based on the simulated flows and the diameter, for each pipe the absolute and relative spare capacity are calculated. Sewsan also allows for storms to be simulated, with the start time, end time, storm peak in mm/h, and expected percentage ingress required as inputs.

2.2.4.3 Hydrograph routing

In Sewsan, hydrographs are routed down the network using time-lag routing. The lag time is calculated using Manning's equation, assuming full flow. The hydrograph shape is retained, but at any point in the network, the individual hydrographs from upstream will be out-of-phase, making the resulting hydrograph appear attenuated (GLS Consulting, 2019).

2.2.4.4 Interpolation and peak capturing

For simulation time steps smaller than one hour, ordinary linear interpolation could result in the actual peak being cut off. Sewsan therefore uses a peak-shifting method to ensure the maximum flow is modelled. This method is conservative and results in hydrographs that slightly over-predict flow volumes (GLS Consulting, 2019).

2.2.5 Cost estimation

Normally, for infrastructure projects, the design is used to compile a bill of quantities, and the cost per item is assigned based on a suitable guideline. Another approach is the use of cost functions, which are normally built as regression models using costing data from previous projects. Cost functions come in many forms, and might account for the total project cost or for a certain subset of costs. For example, Swamee (2001) proposed a cost function where the total installation cost is estimated as the sum of the three cost components presented in Table 2-2. To be reasonably accurate, any costing method should account for the individual sewer network components. The more tailored the costing method is to the locality, the better the estimates are likely to be.

Table 2-2: Cost function example (Swamee, 2001).

Cost Component	Formula	Symbols	
Cost of pipes and jointing per diameter size	$C_p = k_p L D^p$	L D k_p and p	pipeline length diameter regression parameters
Excavation cost per section	$C_e = k_e L(d_1 + d_2)$	L d_1 and d_2 k_e	pipeline length invert depths at link ends regression parameter
Cost per manhole	$C_m = k_m d_m$	d_m k_m	manhole depth regression parameter

2.2.6 Concluding summary

This section provided an overview of the standard sewer design and costing methods used in South Africa, forming the theoretical background to the study. The discussion was mainly applicable to the advanced design stage of a project when most of the project parameters are known. However, the focus of this study is to enable early-stage infrastructure and cost estimates to be made in the absence of detailed design information. Therefore, the remainder of the literature review explores the methods that have been developed, for different reasons, to circumvent this process in order to obtain infrastructure or cost estimates based on limited information. Broadly, this falls into three categories, namely automated generation of entire sewer networks, direct cost estimation, and quantification of the required sewer infrastructure components. Each of these categories is considered in the three sections to follow.

2.3 Automated Generation of a Sewer Network Plan

Automated generation of entire sewer network plans is a topic that has received much attention in recent years. Indeed, it is quite possible that the future of sewer network design lies in this direction. There have been numerous approaches to this concept, and these attempts can be grouped into three categories based on their intended applications, namely: generating the most likely real network for a specific location, optimising the cost-efficient design of a sewer network, and generating random networks for use as case studies. In the following three sub-sections, the work in each of these categories is discussed, although it is the first category that is the most relevant to this study.

2.3.1 Generating most likely real network for a specific location

Blumensaat et al. (2012) developed a tool for generating a realistic hydraulic model of a combined-type sewer for a specific area, using minimal data. By assuming that the sewer network corresponds to the street network, then, with the road layout and a digital elevation model (DEM) as inputs, the sewer layout is generated. A surface flow accumulation algorithm generates the design flows, which are used to size the pipes. The output is a hydraulic model. The model-generating tool was tested on three real-life catchments, and the resulting models were extracted to the EPA-SWMM modelling platform to perform hydraulic simulations. The model was deemed capable of sufficiently imitating the original network layout, length and discharge rates; although the pipe sizing function needed improvement.

A similar model was developed by Greene et al. (1999) to design a complete sewer network with minimal user intervention. As input, the tool requires a GIS map and a topographical model for the service area, and the user must stipulate the locations of manholes. The topography, surface features and street network are then analysed to delineate sub-catchments, generate the pipe network between manholes, and place pump stations and rising mains. When applied to two areas of a town, the authors concluded the resulting network was objectively better than the existing design.

2.3.2 Generating the optimal cost-efficient network design

In order to consider all possible options for a sewer network, and select the optimal cost-effective design, computing power beyond the capabilities of a human is required. Developing algorithms for the optimisation of sewer networks has gained much attention as a research topic. According to De Villiers et al. (2018), this typically consists of two sub-problems – determining the layout of the network elements (nodes and pipes), and then determining the hydraulic parameters of these elements (such as slopes, pipe diameters and materials). Historically, algorithms have been aimed at optimising the hydraulic parameters for a fixed layout. But true optimisation requires both sub-problems to be solved simultaneously. De Villiers et al. (2018) described three approaches to this, namely complete enumeration, separated design, and simultaneous design. In complete enumeration, all viable layouts are generated and the hydraulic design for each layout is completed, allowing the most optimal solution among them to be selected. In separated design, the best among all the possible layouts is selected first, and the optimal hydraulic design is completed for this layout only. This approach may allow the true optimal design to be missed. And finally, in simultaneous design, both sub-problems are optimised concurrently using sophisticated approaches such as ant colony algorithms. Numerous studies have been completed using each of these three approaches. However, this field of research is still focussed more on the optimisation of the algorithms themselves than on their real-world application, therefore it lies beyond the scope of this study.

2.3.3 Generating virtual case studies

In the study of urban drainage, virtual sewer networks are frequently used as case studies to test new methods and software (Urich, et al., 2010; Sitzenfrei, et al., 2010b). This approach has been utilised in cases where there has been an absence of real case studies, or where a very large number of case studies would be required to reach broad conclusions. Several algorithms have been developed for generating virtual sewer networks.

Möderl et al. (2009) developed a tool called ‘Case Study Generator’, which stochastically generates virtual combined-type sewer networks, given certain boundary conditions as input (including the drainage-system length and catchment slope). Sewer lines are generated using a branching process, and assigned slopes and diameters using DEM and rational-method stormwater inputs. The resulting virtual networks can be extracted as models to hydraulic

simulation software. According to Sitzenfrei et al. (2010b), since this tool is based on an oversimplified network system, it does not accurately resemble real-world case study data.

Ghosh et al. (2006) presented the 'Artificial Network Generator' (ANGel), a public-domain application that creates artificial sewer networks with similar geometric characteristics to real networks. It can be applied on a real drainage area with existing land use data, and the sewer lines are generated using a fractal algorithm. It can also work by densifying the layout of an existing or partial network structure. The resulting virtual networks can be extracted as models to hydraulic simulation software. Although the tool can generate artificial networks at the desired drainage densities, the hydrologic results were found to be unreliable.

Sitzenfrei et al. (2010a) developed an urban-planning algorithm called 'Virtual Infrastructure Benchmarking' (VIBe) for the generation entire virtual cities. The parameters of the virtual case studies (for example, land use distribution and population densities) are stochastically varied in realistic ranges to produce numerous case studies with associated elevation maps, land uses, and population distributions. This algorithm was extended with a module to develop sewer infrastructure for each virtual urban case study (Urich, et al., 2010). As such, VIBe generates the input files for the sewer network generation module.

The VIBe algorithm was then further developed by introducing the functionality to dynamically model changes that affect the urban structure over time (such as changing population size, land use and legal standards). This advanced algorithm was named 'Dynamic Virtual Infrastructure Benchmarking' or DynaVIBe (Sitzenfrei, et al., 2010b). DynaVIBe could be a powerful tool for generating dynamic virtual sewer case studies for a given simulation period. The model input parameters can also be tailored to predict a realistic range of the infrastructure requirements of a specific area, as well as to account for specific future scenarios.

2.3.4 Concluding summary

Tools for the automatic generation of sewer network plans show great potential regarding the estimation of sewer network infrastructure and early-stage costing. Of the automatic network generators discussed, two of them might already be applicable to the purposes of this study, namely VIBe and DynaVIBe. However, no examples or investigations of such a utilisation were found.

While it appears that research into automated network generation is driving the future of sewer network design, for now most of these methods are not yet fully developed. Therefore, a simpler approach of statistically estimating the expected costs or infrastructure associated with a network holds much appeal. Examples of such approaches are investigated in the following two sections.

2.4 Direct Capital Cost Estimation of a Sewer Network

In this section, methods are investigated for directly estimating sewer project costs in cases where the infrastructure components are not yet known. In the methods discussed, sewer project cost estimates are enabled using only basic characteristics of the service zone, such as population size, total flow production, area size, and the land characteristics.

2.4.1 Local guideline

A cost benchmarking guide for water services was produced by the South African Department of Water and Sanitation or DWS (PULA, 2016). This document provides the typical unit costs of water services projects and their individual infrastructure components. The costs were derived from the DWS rural water supply projects completed after 1994, as well as from as-built project information from numerous engineering consulting firms and material suppliers. The cost benchmarking guide is intended to aid in decision-making at a local authority, provincial and national level.

Of particular interest to this study are the estimates for the sewer pipeline capital-, operation- and maintenance costs for a level of service where all households have a flush toilet. Table 2-3 presents the capital cost estimations. For different population sizes, the expected length of sewer pipeline is provided. Then, the capital cost of installing this length of pipeline for three different materials is estimated (excluding fees, VAT, preliminary costs, and general costs). It is left to the user to estimate what portion of the network will be made up of each material. The cost estimates are provided per scheme, per capita, and per household. Furthermore, for site-specific considerations such as project size, remoteness, topography, land clearing, contractor availability, geology and land acquisition, adjustment factors were developed to increase or decrease the cost relative to the general case. These are provided in Table A-1 in Appendix A.

Table 2-3: Capital cost for full waterborne sanitation schemes (PULA, 2016).

Scheme Size		Very small	Small	Medium	Large
Number of People		1000	5000	20000	50000
Length of Sewer Pipeline (km)		5	8	17	32
Capital Cost per Scheme	Concrete	R330 919	R1 027 846	R6 151 467	R23 370 596
	uPVC	R628 628	R1 280 549	R5 063 513	R19 002 737
	Lined Steel	R1 279 443	R2 967 559	R11 258 399	R31 397 724
	Average	R746 330	R1 758 651	R7 491 126	R24 590 352
Capital Cost per Capita	Concrete	R331	R206	R308	R467
	uPVC	R629	R256	R253	R380
	Lined Steel	R1 279	R594	R563	R628
	Average	R746	R352	R375	R492
Capital Cost per Household*	Concrete	R1 655	R1 028	R1 538	R2 337
	uPVC	R3 145	R1 281	R1 266	R1 900
	Lined Steel	R6 395	R2 968	R2 815	R3 140
	Average	R3 732	R1 759	R1 873	R2 459

* 5 people per household assumed.

2.4.2 Cost models

Another direct costing approach is a cost model, which expresses the sewer cost mathematically as a function of certain variables. This approach is similar to the traditional cost functions described in Section 2.2.5. However, instead of using detailed design parameters such as pipe lengths, diameters, materials, and depths, the independent variables are basic characteristics of the service zone such as population size, total flow production, area size, dwelling density, and land use. Many such cost models have been developed internationally. An example is the model developed by Balaji et al. (2015). Using data from completed wastewater schemes in 31 towns in India, a regression analysis was performed to determine empirical equations relating the total installation cost (for materials, equipment, and labour) to the population size. Five possible relationship forms were tested, namely: linear, exponential, logarithmic, polynomial, and power. The exponential model performed the best, and the resulting model is presented in Equation 2-11 (Balaji, et al., 2015). Ideally, a cost model should be developed from local data to be considered applicable for a certain wastewater project.

$$C_T = 2.3599 \times 10^5 \times P^{0.7054} \quad \text{2-11}$$

Where

C_T total installation cost (Rupees, 2011-12 market rates)

P population size

2.4.3 Concluding summary

The advantage of direct costing methods is that minimal information and time are required to obtain cost estimates. If the cost estimate can be tailored to project-specific conditions, as in the DWS cost benchmark, then direct cost methods can provide the simplest reliable solution to early-stage cost estimation. However, the ability to predict the required sewer infrastructure components before obtaining an answer that is only related to cost is valuable for several reasons, as mentioned in the Introduction chapter. Consequently, there have been numerous attempts to estimate the sewer pipeline infrastructure for a wastewater scheme using limited information. These are discussed in detail in the next section.

2.5 Estimation of the Infrastructure Components of a Sewer Network

In this section, various methods for estimating the pipeline infrastructure components of a sewer network are investigated. Most of the methods discussed concern estimation of the total pipeline length only; but some do make provision for the diameter distribution. The literature is broadly grouped based on the characteristics that are used as predictor variables. The first sub-section considers methods that use existing urban surface infrastructure as the predictor, and the second sub-section considers methods that use basic population and physical area characteristics of the service zone as predictors. Additionally, the related study field of the hydrological geomorphometry is briefly discussed, insofar as its similarities with and potential application in this study are concerned.

2.5.1 Predictions from existing urban infrastructure

Some studies have sought to quantify existing sewer infrastructure based on urban surface information. Such methods may find use in applications where the existing sewer infrastructure is unknown, lost, or confidential, for example, in asset management or in the valuation of damages from natural disasters.

Haile (2009) investigated the possibility of estimating the characteristics of a combined-type sewer system using urban surface information. Using GIS images, digital elevation models (DEM) and sewer layout plans for nine sub-catchments in a German city, three different possible surface-sewer relationships were examined. The correlation was analysed between the street and sewer network layout, between building sizes and sewer pipe properties, and between surface slopes and underlying pipeline slopes. It was concluded that a) the degree of match between the street layout and the sewer layout varied based on the street pattern style, b) the size of buildings did not predict the properties of the related sewers, and c) the surface slope only predicted the sewer pipeline slopes in limited cases. While these results suggested some correlation between the characteristics of interest, it was concluded that additional data would be required to predict the sewer network characteristics with any reasonable accuracy.

Using GIS data, Kobayashi et al. (2011) attempted to establish whether the length of the road network could be used to estimate the length of the pipelines for the water supply, sewer and low-pressure gas systems in Japanese cities. It was found that the correlation between the lengths of the roads and the water supply pipelines was strong, but only in the densely inhabited districts. Furthermore, it was found that the water pipeline length correlated well with the lengths of the sewer and low-pressure gas pipelines, but with large residuals. It was concluded that in the absence of a better method, this method could be used to obtain rough estimates of the lengths of the water supply, sewer, and low-pressure gas pipelines in densely-inhabited city regions.

Overall, predicting the underlying sewer pipeline properties using urban surface data does show potential to be a fairly reliable approach, but further development in this field would be required. Naturally, such methods are limited in that they are applicable only to existing service zones.

2.5.2 Predictions from basic population and area characteristics

This section focusses on methods for estimating the total pipeline length, or length per diameter category, using service zone characteristics that are available before the detailed design phase. Six such methods, summarised in Table 2-4, are discussed in the following sub-sections. The literature discussed in this section is the most similar to the current study since a common goal is shared, and emphasis is therefore placed on the methodology and notable findings.

Table 2-4: Summary of existing sewer pipeline infrastructure estimation tools.

Method	Predicted Infrastructure	Predictor Variables
DHS, 2019	Reticulation pipeline length per stand.	Land use, Stand size.
Heaney, et al., 1999 A	Reticulation and large-diameter pipeline length per stand.	Dwelling density, Population size.
Heaney, et al., 1999 B	Total pipeline length per diameter category.	Population size.
Pauliuk, et al., 2014	Total pipeline length per diameter for a city.	Population size, Service area length.
Maurer, et al., 2013	Total combined-sewer pipeline length per diameter in a uniform settlement.	Area size, Dwelling density, Flow produced per capita.
Grotepass, 2020	Total water supply pipeline length per diameter.	Peak hour demand, Area size, Land use and topography.

2.5.2.1 Reticulation pipe length per stand, based on land use and stand size (DHS, 2019)

For use in the estimation of groundwater infiltration volumes, the DHS (2019) design guideline provides estimates of the typical reticulation pipe length per stand, for different land uses and stand sizes. The typical reticulation pipe lengths were not obtained through empirical methods but were calculated geometrically, by assuming square plots and a uniform reticulation layout. The results are presented in Table 2-5.

Table 2-5: Typical sewer reticulation length per stand (DHS, 2019).

Land Use		Stand Size (m^2)	Pipe Length (m)
Residential stands	High density, small sized	400 – 670	10 – 13
	Medium density, medium sized	670 – 1 000	13 – 16
	Low density, large sized	1 000 – 1 600	16 – 20
	Very low density, very large sized	1 600 – 2 670	20 – 26
Stands for low income housing	High density, small sized	270 – 400	8 – 10
	Medium density, medium sized	400 – 670	10 – 13
	Low density, large sized	670 – 1 000	13 – 16
Group/ cluster housing	High density	130 – 200	6 – 7
	Medium density	200 – 270	7 – 8
	Low density	270 – 400	8 – 10
Flats	Very high density	80 – 100	4 – 5
	High density	100 – 130	5 – 6
	Medium density	130 – 160	6
	Low density	160 – 200	6 – 7
Agricultural holdings (domestic and irrigation)		< 2 670	> 26
Golf estate (excluding golf course requirements)		< 2 670	> 26
Retirement village		400 – 670	10 – 13

2.5.2.2 Reticulation and large-diameter pipeline length per stand, based on dwelling-unit density and population size (Heaney, et al., 1999)

Heaney (1999) developed a tool to estimate the total sewer pipeline length required for a service zone based on the number of dwelling units. The reticulation length per stand (or 'lot pipe') was calculated in a similar manner to the DHS tool, by calculating the average stand size using the dwelling density, and then calculating the pipeline length per stand geometrically. Following that, the typical additional lengths of large-diameter collector and bulk pipes per stand were generated using empirical data, for different population sizes. The final pipeline length estimation tool is presented in Table 2-6.

Table 2-6: Estimated pipeline length per dwelling in a sewer scheme (Heaney, et al., 1999).

Dwelling-unit Density (units/acre)	Lot Pipe (ft/unit)	Added Larger Pipe for Various Population Sizes (ft/unit)		
		1 000	10 000	100 000
2	70	10.5	14	28
4	41	6.15	8.2	16.4
6	31	4.65	6.2	12.4
8	21	3.15	4.2	8.4
10	11	1.65	2.2	4.4

2.5.2.3 Pipeline length per diameter, based on population size (Heaney, et al., 1999)

Additionally, Heaney et al. (1999) provided a second tool that could be used to estimate the pipeline length within different diameter categories for a sewer scheme, based on the population size. The results are displayed in Table 2-7. This tool was developed by summarising data compiled by Dames & Moore (1978) in a study of 455 sanitary sewer construction projects in the USA.

Table 2-7: Sanitary sewer pipe based on city size (Heaney, et al., 1999).

Population Range	Length of Various Pipe Sizes (km)					% Small Pipes*
	<8''	8'' – 14''	15'' – 24''	>24''	Total	
> 500 000	1.761	63.809	24.093	20.352	110.014	0.60
250 000 – 500 000	7.821	42.041	11.941	8.031	69.834	0.71
100 000 – 250 000	8.063	56.044	9.112	7.419	80.638	0.79
50 000 – 100 000	16.192	48.159	9.830	8.427	82.607	0.78
25 000 – 50 000	14.859	55.698	10.861	5.475	86.893	0.81
10 000 – 25 000	30.643	77.161	11.690	3.570	123.065	0.88
2 500 – 10 000	38.603	119.505	20.503	6.095	184.706	0.85

* Small pipes defined as having diameter < 14''.

2.5.2.4 Pipeline length per diameter for a city, based on population and area size (Pauliuk, et al., 2014)

Pauliuk et al. (2014) developed a model to estimate the total pipeline length per diameter of a city-scale sewer network, based on the population and area size. A regression model of the form $L = \alpha P^{\beta} {L_0}^{\gamma}$ was developed, where L is the total pipeline length in kilometres, P is the population size, and L_0 is the length of a diagonal drawn across the city area. The regression constants α, β, and γ were determined using nearly 100 data points, from 35 settlements and cities on five continents, ranging from 500 to 23,000,000 inhabitants. The final model form is presented in Equation 2-12. However, this model could only account for 75% of the sample variance, and it was recommended the total wastewater volume, dwelling density, and topography should also be factored in. The associated diameter distribution was determined as the average length per diameter of sewer pipes in three Norwegian cities. It is noted that this tool did not distinguish between combined and separate sewers.

$$L = 0.3P^{0.55} {L_0}^{0.51} \qquad 2\text{-}12$$

2.5.2.5 Combined-type sewer pipeline length per diameter, based on area size, dwelling density, and wastewater production per capita (Maurer, et al., 2013)

A model was developed by Maurer et al. (2013) to estimate the combined-type sewer pipeline length per diameter for a settlement based on area size, dwelling-unit density, and wastewater production per capita. It was assumed that an appropriate settlement should have fairly homogeneous plot sizes, a uniform population density, continuously-falling gradients, a compact shape, and a consistent rainfall pattern. The network geometry was then approximated as illustrated in Figure 2-2. The model development process comprised three steps:

- Setting up expressions for the total pipeline length and diameter distribution based on geometric and hydraulic calculations.
- Performing a sensitivity analysis to determine which variables were significant and which could be modelled as constants.
- Performing a regression analysis to determine the values of statistical constants in the equations, using data from 586 Swiss towns and cities.

The eight candidate variables in the first expressions included settlement area size, population size, dwelling-unit density, wastewater production per capita, surface runoff area, surface runoff coefficient, and the pipe internal roughness. Through the sensitivity analysis, it was found that most of the variance in the results was accounted for by only three of the variables. The final models (not provided in the literature source) therefore expressed the total pipeline length and diameter distribution as a function of the area size, dwelling-unit density, and wastewater production per capita. It was concluded that the model could emulate the principal characteristics of a sewer system, although the accuracy of estimation was not quantified. Additional findings from this study were that the pipeline infrastructure was highly dependent on the settlement layout, and that larger areas with higher populations had a higher proportion of large pipes.

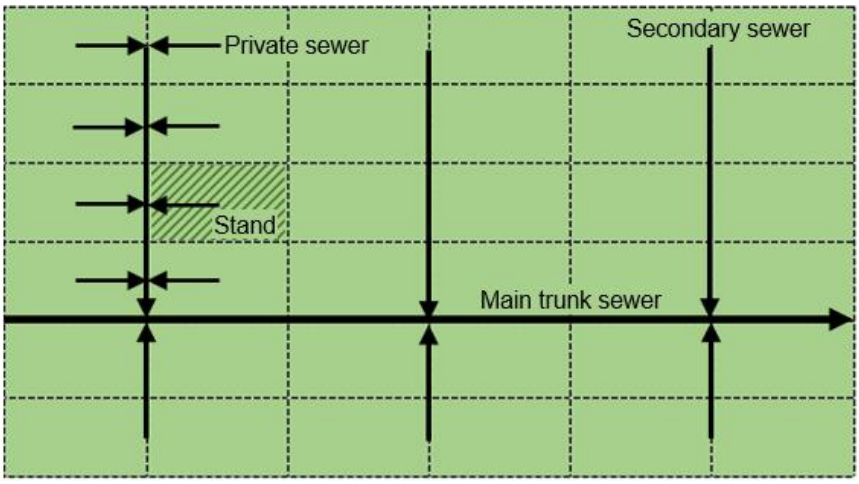


Figure 2-2: Assumed sewer layout and pipeline hierarchy (Maurer, et al., 2013).

2.5.2.6 Total water supply pipeline length per diameter, based on peak hour demand, area size, land use and terrain category (Grotepass, 2020)

Using South African water supply network data, Grotepass (2020) developed a linear regression model expressing the total peak hour demand of a water supply network as a function of certain predictor variables. The candidate predictor variables used to develop the model included total pipeline length, total pipeline volume, land area, area shape factor, terrain category, reservoir distance from area centroid, and reservoir elevation above mean terrain elevation. The resulting model is presented in Equation 2-13, where the peak hour demand (L/s) is a function of the total pipeline length (m), the total service zone area (hectares), and a diameter distribution factor SF. SF is dependent on the land use, area size and terrain characteristics of the service zone.

Equation 2-13 can be rearranged to form Equation 2-14, thus providing the total pipeline length as a function of the peak hour demand, area size, land use, and terrain category. Pipeline diameter distributions were also generated for different types of networks, to enable the total pipeline length to be disaggregated into pipeline lengths per diameter.

$$Peak\ hour\ demand = 9.85 + SF \times Total\ pipeline\ length - 0.0725 \times Area \quad 2\text{-}13$$

$$Total\ pipeline\ length = \frac{1}{SF}(Peak\ hour\ demand + 0.0725 \times Area - 9.85) \quad 2\text{-}14$$

The study by Grotepass (2020) is acknowledged as the incentive for the current study, which attempts to establish an analogous relationship for sewer networks, following a similar methodology.

2.5.3 Hydrological geomorphometry: relating river systems to catchment form

Similarly to how the literature in Sections 2.5.1 and 2.5.2 was investigated with the intention of identifying potentially-useful outcomes that could be applied in the current study, in this section, hydrological geomorphometry is briefly considered as a related field of study offering a body of knowledge with potential application in this study.

Hydrological geomorphometry is the study of how the shape and topography characteristics of a river catchment area relate to the qualities of the resulting river system. It is fundamental to this field that the length and volume of a stream are directly related to the morphometric (or 'form') characteristics of the catchment. As such, there are many well-established laws describing the relationships between catchment form and river characteristics (Zavoianu, 1978). In this sense, there are significant similarities between hydrological geomorphometry and the aims of this study. That is, this study aims to quantify the relationship between the form of a service zone and the characteristics of the associated sewer network. While the laws of hydrological geomorphometry may be true for natural river catchments, this does not imply that the same is true for man-made sewer 'catchments'. However, the variety of methods for quantifying the form characteristics of river catchments found in hydrological geomorphometry do show potential to be applied successfully to sewer 'catchments' to allow better representation of their form.

Consequently, indicators originating in the field of hydrological geomorphometry were included in a thorough investigation into methods for quantifying service zone form characteristics (area size, shape, and topography) provided in Appendix B. The investigation in Appendix B provides supplementary research for Chapter 4 on Data Collection, therefore the reader is referred to it at a later stage in the report.

2.5.4 Concluding summary

It is clear that attempting to estimate the pipeline infrastructure of a sewer (or even water supply) network based on limited information is not a new concept. To this end a variety of different approaches were found in the literature. However, no methods were found that were developed using South African or even African data on a scale that could be applicable to future developments. In this sense, the preceding discussion highlights the potential usefulness of such an estimation tool.

The findings from the various studies also helped to identify important service zone characteristics for consideration as candidate variables in the current study. These were the land use or network layout, population size, area size, dwelling-unit density or stand size, wastewater flow production per capita, total flow production, and topography. Furthermore, the field of hydrological geomorphometry offered a number of methods for quantifying catchment form characteristics, which were investigated in this study in order to identify suitable methods for quantifying the form characteristics of service zones.

2.6 Literature Review Concluding Summary

The two major aims of the literature review were to orientate the proposed study relative to the surrounding literature to clarify its purpose and potential benefit; and to identify findings from similar studies that could be accounted for in the proposed study.

Relating to the first aim, several conclusions were made. Firstly, while there has been significant progress towards the automated generation of entire sewer networks, this approach is not yet accessible for use on a project level, and there is still a need for simpler cost and infrastructure estimation methods. In terms of direct cost estimation, there is a local guideline available for this purpose, but this guideline would be complemented by better estimates of the sewer pipeline

infrastructure. Therefore, a South African tool for estimating sewer pipeline infrastructure based on early-stage information could still offer considerable benefit. While many similar tools developed using varied approaches were found in the literature, none were found that could be reliably applied in South Africa. Furthermore, it is expected that much of the future development of the South African sewer network will take place in low income and rural areas, which should be accounted for.

Relating to the second aim, the literature review helped to identify several service zone characteristics that should be considered as potentially influential in the proposed study, namely land use or network layout, population size, area size, dwelling-unit density or stand size, wastewater flow production per capita, total flow production, and topography. Additionally, a number of catchment form indicators were identified from an investigation of the related study field of hydrological geomorphometry, which could aid in the accurate quantification of certain service zone characteristics.

With the relevant literature having been investigated in detail, and the important outcomes highlighted, the rest of this report now focusses on the current study. Chapter 3, the Research Design, provides an overview of the methodological approach followed in this study.

Chapter 3
RESEARCH DESIGN

In order to achieve the main aim of this study, to develop a method for estimating the sewer pipeline infrastructure of a service zone using limited information, three major study outcomes were identified. Study Outcome I required a model for estimating the total pipeline length; Study Outcome II required the diameter distribution of a typical sewer network; and Study Outcome III required the expected number of manholes per kilometre of pipeline. This chapter briefly outlines the methodological approach taken to realise the stated study outcomes.

The approach taken to realise the stated study outcomes was a statistical one. This necessitated two major methodological components, namely data collection and statistical analysis. For the data collection component, a suitable and sufficient data source characterising a large number of service zones and associated sewer networks had to be identified. From this, an appropriate dataset of sample networks had to be extracted to be used for statistical analysis.

For the statistical analysis component, each study outcome necessitated a unique statistical approach. For Study Outcome I, the chosen method was to develop a regression model to express the total pipeline length as a function of a combination of physical characteristics of the service zone. This method was chosen to enable precise estimation of the total pipeline length, as well as to allow the relationship between the total pipeline length and the service zone characteristics to be quantified and understood.

Regarding Study Outcome II, the distribution of pipe diameters does not lend itself to precise statistical estimation, since pipe diameters are dependent on specific factors, such as the network layout or individual pipe slopes. Therefore, a simple and practical solution of finding the average diameter distribution for similar networks was chosen. This pragmatic approach required similar networks to be identified based on known service zone characteristics, such as land use, area size, or population density. In order to increase the viability of the chosen method, the categories and category boundaries had to be set based on logical consideration, such that meaningful differences between the distributions would be obtained.

Lastly, for Study Outcome III, a simple average of the number of manholes per kilometre of pipeline was required. However, it was expected that this manhole frequency could be influenced by certain service zone characteristics. Therefore, the approach taken was to establish first whether the manhole frequency was influenced by any service zone characteristics, and then, based on the outcome, to determine the expected number of manholes per kilometre of pipeline in such a way that any influences would be accounted for.

The methodology is described in detail in the five chapters to follow. Chapter 4 details the data collection process, Chapter 5 provides an overview of all the relevant regression methods, and then the statistical analysis procedures for Study Outcomes I, II and III are detailed in Chapter 6, Chapter 7, and Chapter 8, respectively.

Chapter 4
DATA COLLECTION

This chapter provides a comprehensive report of the data collection process. The aspects covered in this chapter include the raw data source, sample network definition and selection, modifications to the source data, quantification of the network characteristics, data screening, and the final dataset. The data collection is described in detail, since the applicability of the final infrastructure estimation tool is dependent on the data used to develop it. Where necessary, technical aspects of the data collection process are described and discussed in the appendices, to ensure that the study is repeatable, that the correct model inputs can be calculated by future users, and that the logic behind important methodological decisions is clear.

4.1 Raw Data Source

Data in the form of computer models of various South African sewer networks were provided by GLS Consulting (GLS), an engineering consulting firm based in Stellenbosch, South Africa. GLS specialises in the analysis, planning and management of municipal networks for water supply, sanitation, stormwater, and electricity. GLS also provides software products for the design and analysis of such networks, including the Sewsan software for sewer system analysis. In terms of waterborne sanitation systems, GLS provides services such as modelling existing networks, optimising network design and operation, and developing long-term master plans for upgrades to the system. The clients of GLS include major South African municipalities, such as City of Cape Town, City of Tshwane, Ekurhuleni, Johannesburg Water, George, and Buffalo City, as well as numerous other municipalities throughout the country and internationally. GLS was therefore considered a highly credible and reliable data source.

The data provided by GLS consisted of working models of entire sanitary sewer infrastructure networks for five South African municipalities, housed in the Sewsan software. The data included two major metropolitan municipalities and three local municipalities, with an amassed total of 20 660 km of gravity pipeline. The models represented areas with varied characteristics, such as historic and new, flat and hilly, and urban and rural. Therefore, the models were considered to be representative of the variety of conditions present in typical South African sewer networks.

Each sewer model contained comprehensive data about the network. All hydraulically-relevant elements of the networks and their associated characteristics were included in the models. This included gravity pipes, rising mains, manholes, diversions, other special structures, pump stations, and terminal structures such as wastewater treatment plants (WWTPs) and conservancy tanks. Each municipality had its own calibrated set of unit hydrographs for each land use. The number of unit hydrographs per land use were assigned at the relevant manholes, as well as the AADD which had been regularly updated with water meter data, obtained in electronic format from the municipal treasury systems. The simulation results detailed the relevant hydraulic conditions in every pipe and structure. The models were embedded in a GIS-environment, allowing all elements to be located in a geospatial coordinate system, overlaid on a high-resolution aerial photograph background, and linked to the digital elevation model (DEM) providing the in-situ ground elevations.

In summary, the data was considered reliable, representative, and comprehensive. However, these very large models had to be broken up into finite networks before the necessary information could be extracted for analysis.

4.2 Definition of a Sample Sewer Network

Each municipality-scale model comprised several discrete large networks draining to individual WWTPs. In turn, each discrete large network was made up of converging sub-networks. Since the envisioned infrastructure estimation tool was intended for development-scale networks, and since a development-scale network would typically represent a sub-section of a larger network only, the sample networks used for this study had to be singled out from the existing larger networks. In order to select sample networks of an appropriate scale, a ‘sample network’ had to be defined first in the context of the study.

Firstly, considering a sample network as a sub-network of a larger system, implies a connector pipe between the convergence point of the sub-network (point A) and the point where it flows into the larger system (point B). The connector length is dependent on the distance between points A and B. The connector length could probably be roughly calculated for a specific service zone with more accuracy than could be expected from a statistical prediction. Therefore, the connector was excluded from the definition of the network. The end point of a network was then defined as the first point receiving the full combined flow from the service zone (point A).

Another important consideration was the presence of rising mains. A rising main is required when the endpoint of a sub-network is lower in elevation than the pipe into which it must flow. A rising main was considered too complex a factor to predict statistically using basic characteristics of the service zone. For example, an area might be hilly, but this does not provide enough information as to whether rising mains will be required. Rising mains were therefore excluded from the definition of a network. Nonetheless, the tool is still theoretically applicable to sub-catchments of a larger service zone, and the length of rising mains between sub-catchments can be estimated.

In summary, a sample network was defined as a system of gravity sewer pipes converging at a single endpoint at the first manhole that receives the full combined flow from the service zone.

4.3 Sample Selection

The definition in Section 4.2 still left much room for interpretation when selecting sample networks. Figure 4-1 provides an example of how a sample network singled out from a larger network could be chosen in different ways. In Option 1, the whole network section is captured as the sample network; in Options 2 and 3, only a smaller portion of the network section is captured as the sample network.

Figure 4-1: Example of different options for capturing a network section.

In order to guide the sample selection process, some additional requirements for the sample networks were identified. It was required that the sample networks should span a size range typical of realistic new developments, contain sufficient samples for each land use category, and be varied in characteristics such as shape, pipe layout, dwelling density, and topography.

To ensure the sample network requirements listed above were met, a master planning database containing basic zoning information for 900 planned future developments was provided by GLS. Statistics from the future development data were summarised as presented in Table 4-1. The area size distribution served as a guide to ensure that a realistic range of area sizes was accounted for in the collected dataset. The percentage of properties per land use land use helped to ensure that each land use was appropriately represented in the collected dataset. The statistics of the collected dataset were checked regularly against Table 4-1 during the sample selection process to ensure that the dataset was reasonably reflective of reality.

Table 4-1: Area size and land use distribution of the future developments database.

Land Use Category *	Area Percentile (ha)					% Properties
	Min	25th	50th	75th	Max	
General Residential	0	6	19	52	1076	57
Low Income Residential	0	5	22	60	328	10
Non-Residential	1	7	17	42	807	31
Large	4	12	29	38	139	2
All	**0**	**6**	**19**	**50**	**1076**	**100**

* Land use categories are defined in Section 4.5.

In order to ensure the sample set had good variation in terms of the other network characteristics such as shape and topography, selection of the networks was done as randomly as possible. However, continuous visual assessment of the selected networks was used to ensure that certain conditions were not seriously over- or under-represented. For example, most of the networks were in fairly flat areas, so the few service zones near visible ridgelines or mountains were actively sought out to ensure steeper terrain was sufficiently represented.

Despite the measures taken, model selection remained a subjective process. To reduce the potential impact of this subjectivity, as well as to ensure that sufficient samples with different combinations of features were obtained, a very large dataset of 500 networks was collected.

4.4 Modifications to Sample Networks

While selecting the data points, two types of changes were made to the source data using Sewsan's 'planning analysis' function (see Section 2.2.4.1) for resizing pipes. Firstly, since all networks are required to accommodate a spare capacity of at least 30%, it was considered that the source data should also represent networks operating under these conditions. However, not all pipes in the existing systems satisfied this constraint. The GLS team recommended that a planning analysis should be run on the sample networks to ensure that all pipes were operating with adequate spare capacity. The majority of pipes had ample spare capacity, and only a few larger pipes were resized, resulting in a small overall impact on the data.

The second case where the pipes were resized came about from the cutting of sample networks out of larger networks. In most cases, sample networks could be isolated from the main network by deleting the connection immediately downstream of the sample network endpoint. But practically, to obtain a large enough dataset, sometimes lines conveying flow generated from upstream sub-networks outside of the service zone of interest had to be cut off to isolate a desired sample network, as illustrated in Figure 4-2. The pipes in the desired sample network downstream of the cut-off were often larger due to higher flows upstream. Therefore, all pipes directly downstream of the cut-off were resized to represent the realistic diameters that would be expected in the absence of the external flow from the upstream network. The technicalities of implementing these modifications are described in detail in B.7.

Of course, changes to the source data put the validity of the results at risk. Care was taken to ensure that the results would not be invalidated by the changes. In the modification process, no substantial changes were made to pipe lengths, therefore, the pipeline length estimation models (Study Outcome I) were not affected. Only the diameter distribution (Study Outcome II) could have been affected, but care was taken to ensure that the resulting diameters were realistic representations of how real networks would look.

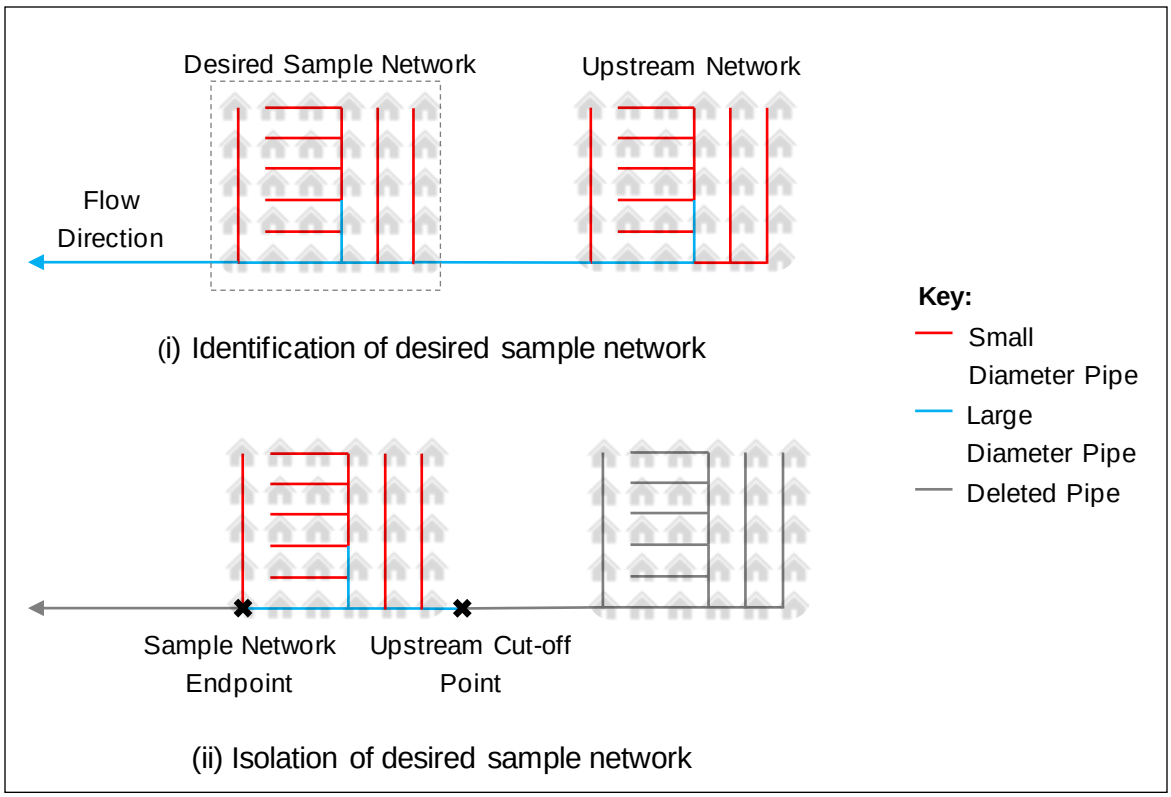


Figure 4-2: Cutting a sample network out of a larger network.

4.5 Quantification of Sample Network Characteristics

Based on the sewer network design principles, the findings from the literature review, and the model requirements, the characteristics of interest that had to be quantified for each sample network were identified as: total pipeline length, diameter distribution, number of manholes, land use, area size, design flow ('flow'), dwelling density, area shape ('shape'), and topography. Before these characteristics of interest could be quantified, certain raw data had to be extracted from each sample network. Table 4-2 presents the raw data extracted from each sample network.

Table 4-2: Raw data extracted from Sewsan models.

Sample Network Data	Description
Lengths and diameters of all individual pipes	A database of all pipes in the sample sewer network, with their associated lengths and diameters.
Manholes and other junction structures	A database of the manholes and all other junction structures (diversions, rodding eyes, top ends, and the occasional flow meter). All junction structures were assumed to have an associated manhole.
Service zone polygon	A polygon drawn along the service zone boundary, with an associated area ($\boldsymbol{A}$), perimeter ($\boldsymbol{P}$), and coordinates of the centroid of its bounding rectangle ($\boldsymbol{X_c, Y_c}$). Appendix D.1 provides further information on how the polygon was defined.
Unit hydrographs	The number of unit hydrographs of each land use type serviced by the network. Appendix D.2 provides further information on how the unit hydrographs were assigned in the source models.
Total user PDDWF	The total PDDWF (kL/d) generated by users that enters the network, excluding infiltration and ingress, calculated using the AADD method. Appendix D.3 provides further information regarding the flow definition and calculation.
Digital elevation model (DEM)	The XYZ coordinates of all DEM points lying within the service zone polygon, using a 25 m DEM grid size. The mean elevation ($\boldsymbol{H_{mean}}$) and the elevations of the lowest ($\boldsymbol{H_{min}}$) and highest ($\boldsymbol{H_{max}}$) DEM points in the service zone were also recorded.
Sample network endpoint	The XYZ coordinates of the sample network's most downstream convergence ($\boldsymbol{X_{mouth}, Y_{mouth}, Z_{mouth}}$).

The raw data described in Table 4-2 was used to quantify the network characteristics of interest, thus forming the variables for the model-building phase. Table 4-3 (continued on the next page) defines these variables, along with the quantification method and characteristic of interest represented by each. The representative variables for the service zone's form characteristics (area size, shape and topography) were selected based on a thorough investigation of methods for quantifying study area form, provided in Appendix B. For some of the characteristics of interest, multiple representative variables were used since it was unclear at the outset which variable characterised that property the best.

Table 4-3: Variables representing network characteristics of interest.

Variable	Network Characteristic	Definition	Reference
Total Pipeline Length per Diameter	Pipeline length; diameter distribution	The sum of all pipe lengths for each unique diameter.	-
Number of Manholes	Manhole distribution	The total number of manholes and other junction structures.	-
Land Use Category	Land use	Land use category in Table 4-4 that best describes the sample network based on percent contribution to total PDDWF[2]. Appendix D.4 provides further information on land use grouping and classification.	-
Area	Area size	The size of the service zone polygon (representing the plane area of the service zone).	-
PDDWF	Flow	The total user flow production in the form of PDDWF (kL/d).	-
Number of Unit Hydrographs	Dwelling density	The total number of unit hydrographs of all land uses serviced by the network.	-
Circularity Ratio	Shape	Formula: $\frac{4\pi A}{P^2}$	(Miller, 1953)
Elongation Ratio	Shape	Formula: $\frac{1.129\sqrt{A}}{L}$; $L = \frac{P}{4} + \sqrt{\left(\frac{P}{4}\right)^2 - A}$ if $A < \left(\frac{P}{4}\right)^2$; $L = 4\left(\frac{A}{P}\right)$ if $A > \left(\frac{P}{4}\right)^2$	(Schumm, 1956)
Centroid-mouth Relative Radius	Shape	Formula: $\frac{\sqrt{(X_c - X_{mouth})^2 + (Y_c - Y_{mouth})^2}}{\sqrt{A}}$	-

[2] The dominant land use category in each sample network, based on the contribution to total PDDWF, generally had a high percentage dominance; therefore, identifying the dominant land use category based on the contribution to total number of UHs instead would probably give the same conclusion in the majority of cases.

Table 4-3 (continued).

Variable	Network Characteristic	Definition	Reference
Mean Slope of Perimeter	Topography	Formula: $\frac{2(H_{max} - H_{mouth})}{P}$	(Zavoianu, 1985)
Mean Slope of Basin	Topography	Formula: $\frac{H_{max} - H_{min}}{L}$; $L = \frac{P}{4} + \sqrt{\left(\frac{P}{4}\right)^2 - A}$ if $A < \left(\frac{P}{4}\right)^2$; $L = 4\left(\frac{A}{P}\right)$ if $A > \left(\frac{P}{4}\right)^2$	(Schumm, 1956)
Melton Ruggedness Number	Topography	Formula: $\frac{H_{max} - H_{min}}{\sqrt{A}}$	(Melton, 1965)
Surface Area Ratio	Topography	Formula: $\frac{Real\ Surface\ Area}{A}$ See Appendix D.5 for calculation of the real surface area from the DEM.	-
Total Relief	Topography	Formula: $H_{max} - H_{min}$	(Zavoianu, 1985)
Mean Relief	Topography	Formula: $H_{mean} - H_{mouth}$	(Wilson & Gallant, 2000)
Elevation Standard Deviation	Topography	The standard deviation of elevations of all DEM points.	-
Ruggedness Number	Topography	Formula: $\frac{Total\ Pipeline\ Length \times (H_{max} - H_{min})}{1000 \times A}$	(Strahler, 1964)
Deviation from Mean Elevation	Topography	Formula: $\frac{H_{mean} - H_{mouth}}{Elevation\ Standard\ Deviation}$	(Wilson & Gallant, 2000)

Table 4-4: Land use categories.

Land Use Category	Land Uses[3]
General Residential	Very high income/ low density residential High income/ medium density residential Medium income/ high density residential Cluster Flats Farm/ agricultural holdings
Low Income Residential	Low income/ very high density residential
Non-Residential	Business/ commercial Educational Government/ institutional Industrial Mixed
Large	Large Public open space

4.6 Data Screening and Pre-Processing

In total, 500 sample networks were collected, and for each one, the values for the 18 variables in Table 4-3 were quantified. Preliminary data screening was then implemented to clean and simplify the dataset. Of the 500 sample networks, 27 were removed due to significant irregularities in the DEM, which reduced the dataset to 473 points. Furthermore, the elongation ratio was removed as a variable since it became inaccurate after a certain critical elongation value, which reduced the number of variables to 17. Lastly, the final network models contained over 600 unique internal pipe diameters. Table 4-5 shows the internal and nominal diameters for standard, locally available pipe sizes, provided by GLS. For practicality, each pipe diameter was rounded up to the nearest internal diameter presented in Table 4-5.

[3] Table 4-4 does not provide an exhaustive list of land uses, rather the ones that were present in the source data models. Any land use not listed in Table 4-4 should be assigned to the land use category to which it most logically belongs.

Table 4-5: Internal and nominal diameters of allowed new pipe sizes.

Diameter (mm)			
Internal	**Nominal**	**Internal**	**Nominal**
104	110	633	675
151	160	704	750
188	200	762	825
235	250	843	900
297	315	1008	1050
335	355	1149	1200
377	400	1290	1350
419	450	1423	1500
488	525	1602	1650
559	600	1717	1800

4.7 Final Dataset

The final dataset consisted of 473 sample networks. The area size and land use distribution of the final collected dataset is displayed in Table 4-6. The values in Table 4-6, particularly the maximum area sizes and number of data points, were subject to change during the analysis due to the removal of outliers and influential points.

Table 4-6: Area size and land use distribution for final dataset.

Land Use Category	**Area Percentile (ha)**					**Data Points**	
	Min	**25th**	**50th**	**75th**	**Max**	**% of Total**	**Number**
General Residential	2	18	45	107	1096	51	240
Low Income Residential	1	11	25	94	774	23	113
Non-Residential	4	20	39	88	445	20	92
Large	3	8	28	116	918	6	28
All	**1**	**18**	**39**	**99**	**1096**	**100**	**473**

4.8 Limitations

A general limitation for the data collection process was that all the results that could be generated using the dataset, were constrained by which data were selected and how. For example, since the sample networks selected represented service zones on a small to large development scale, the results generated using this dataset were also confined to such a scale and would not be applicable for entire towns or cities.

4.9 Ethical Considerations

To satisfy the ethical requirements of the data provider, all relevant municipalities were contacted for consent before the network data was disclosed, and it was requested that the municipalities remain anonymous. In line with the ethical requirements of Stellenbosch University, since the data could not be linked to any individual or client municipality, and since the study was of a non-sensitive nature, no ethical clearance was required.

4.10 Data Collection Concluding Summary

This chapter provided a comprehensive report of the data collection process, with respect to the raw data source, sample network definition, sample selection, modifications to the source data, quantification of the network characteristics of interest, data screening, and the final dataset. In each of these steps, care was taken to ensure that the data samples selected were of high quality, varied, and reflective of reality, thus helping to ensure as far as reasonably possible that the major assumption of this study, that the existing networks used in the analysis were designed to an optimal standard, was met. Where necessary, detailed descriptions of certain technical aspects of the data collection process were provided in the appendices to allow for study repeatability, as well as to ensure that future users of the infrastructure estimation tool would be able to accurately calculate the correct model inputs. The following chapter, Chapter 5, provides a background in regression analysis, with focus on how the regression methods were applied in this study. As such, it serves as a technical guide to the ensuing analysis chapters, namely Chapter 6, Chapter 7, and Chapter 8.

Chapter 5
REGRESSION METHODS

One of the required outcomes of this study, namely Study Outcome I, was the development a set of models to estimate the total pipeline length as a function of the physical characteristics of a service zone. The approach was to develop multivariate linear models using regression analysis. In multivariate linear regression analysis, a dataset is used to generate a model of the form given in Equation 5-1, where y denotes the dependent variable, x_i denotes the independent or predictor variables, and β_i denotes the regression coefficients.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad 5\text{-}1$$

To develop a regression model, the right independent variables x_i in the right forms must be selected, and reliable estimates of the regression coefficients β_i must be generated using regression analysis, such that the model can produce acceptably accurate estimations of the dependent variable y. The field of regression offers an extensive range of methods for model development, from which the right regression methods must be selected, considering the specific analysis requirements and data characteristics.

The purpose of this chapter is twofold. Firstly, it is to provide the reader with sufficient background knowledge of regression to ensure complete understanding of the data analysis detailed in Chapters 6 to 8. Secondly, it is to justify the use of certain regression methods and to describe how they were implemented. Certain fundamental regression principles are explained, but the reader is referred to external sources for more detailed information.

This chapter begins with a background of the most fundamental form of regression analysis, ordinary least squares regression (OLS). This is followed by a detailed discussion of the assumptions that must be met by any OLS model. A variation of simple OLS regression is then described, namely weighted least squares regression (WLS). Thereafter, the important aspects of model development are covered, namely sample size, outliers and influential points, model building, principal component analysis, as well as model evaluation and comparison techniques.

5.1 Ordinary Least Squares Regression (OLS)

The standard approach for regression analysis is to first try ordinary least squares regression (OLS), often simply referred to as linear regression. In a regression model, the errors (or residuals) refer to the difference between the actual (observed) value and the fitted (predicted) value for each observation or data point. In OLS, the regression coefficients (β_i) are estimated such that the sum of the squared errors is at its minimum. OLS is the most common regression method, as it is simple, and it is the most accurate regression method if the assumptions are met. Montgomery and Runger (2014) provides a thorough mathematical explanation of how OLS regression works. However, numerous statistical software packages provide OLS functionality that require only the dataset, and the independent and dependent variables, to be provided as inputs. It is noted that the statistical software packages used for this study were the 'Statsmodels' and 'Scikit-learn' packages embedded in the Python programming language.

While OLS is a linear regression method, it does allow nonlinear terms to be included in the model. A nonlinear term may be included by first applying a nonlinear transformation to the variable of interest, and then representing the transformed variable as a linear variable in the OLS regression. For example, a quadratic relationship between y and x may be modelled as shown in Equation 5-2, by setting $x_1 = x$ and $x_2 = x^2$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad 5\text{-}2$$

5.2 Assumptions of OLS

OLS is subject to certain important assumptions, namely linearity, independence, constant variance, lack of multi-collinearity, and normality. Each of these assumptions is discussed in the following five subsections, in terms of the definition, causes and consequences of violation, checks, and solutions. These assumptions were checked for every model compiled during this study before the results could be taken as valid.

5.2.1 Linearity

The relationship between the dependent and independent variables is assumed to be linear in OLS. If this assumption is violated, then the form of the true underlying relationship has been modelled incorrectly. In this study, several checks were performed to ensure linearity for each regression model, as described in Table 5-1.

Table 5-1: Checks performed for the linearity assumption.

Check	Description
Scatter Plots	Scatter plots of the dependent variable versus each independent variable should be reasonably straight and show no obvious bends.
Residual Plots	Scatter plots of the residuals versus the dependent and each independent variable should be randomly scattered and show no trends.
Partial Regression Plots	Partial regression plots show the relationship between the dependent and a single independent variable after the effects of the other independent variables have been accounted for. For an independent variable of interest, the residuals of the *dependent variable* after a regression on the *other independent variables*, are plotted against the residuals of the *independent variable of interest* after a regression on the *other independent variables* (De Veaux, et al., 2011). Figure 5-1 displays an example of a partial regression plot. The scatter should be reasonably straight with no obvious bends.

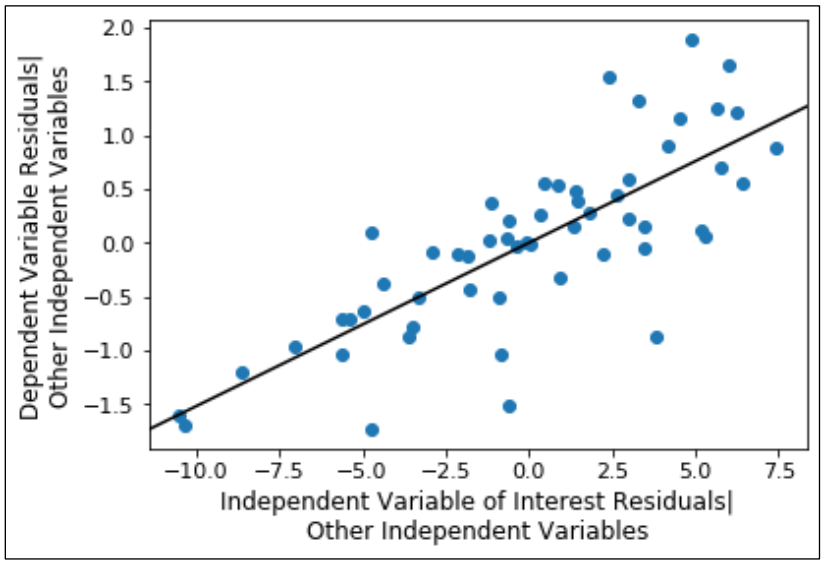


Figure 5-1: Example of a partial regression plot.

If a nonlinear relationship is identified between the dependent variable and an independent variable, a transformation is normally required to model the nonlinear relationship. However, it is not always clear what kind of transformation this should be. A useful tool to determine the transformation required, for simple curves with no inflection points, is the ladder of re-expression developed by Mostellar and Tukey (1977). The ladder or re-expression tool is illustrated in Figure 5-2. The user first identifies the quadrant on the diagram that resembles the curve. The curve suggests whether x or y should be increased or decreased. Transformations can be applied to x or y or both; however, beginning with only x is generally recommended for simplicity. Then, the ladder of re-expression on the right indicates transformations that can potentially correct the curve. $x \rightarrow x$ indicates the starting point. The user tests the different transformations by moving up or down as suggested by the figure, where each step further away from the starting point represents a stronger transformation. The best transformation is identified as the one that visually results in the best fit according to 'check plots' such as those in Table 5-1. However, it must be ensured that the model is not overfitted. Overfitting occurs when a model attempts to account for variation in the data that is in fact part of the random scatter.

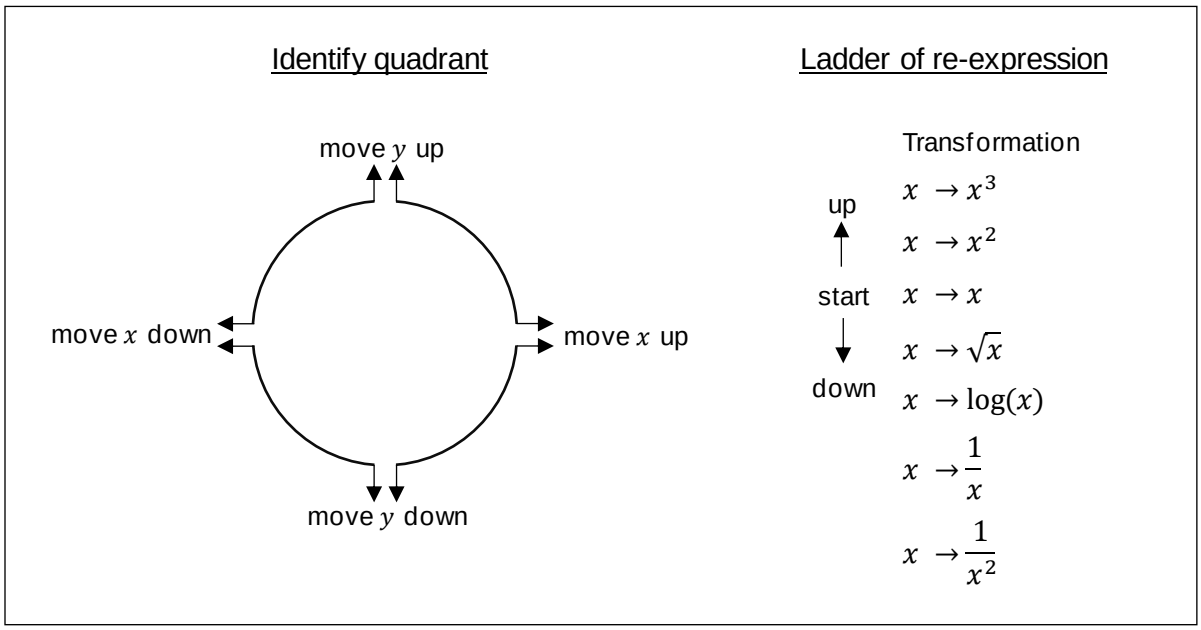


Figure 5-2: Ladder of re-expression (Mostellar & Tukey, 1977).

5.2.2 Independence

The residuals of an OLS model should be independent, which means that their size should be unrelated to their order of observation. This is mainly a concern in time-related data. If the residuals are not independent, this might indicate that a meaningful factor has not been accounted for in the model, or that the measurement accuracy changed over the course of the data collection process. If the independence assumption is violated, the cause must be identified and rectified accordingly, which might require the data collection to be revisited. The check for independence used in this study is described in Table 5-2.

Table 5-2: Check performed for the independence assumption.

Check	Description
Scatter Plot of Residuals versus Observation Order	A scatter plot of the model residuals versus their order of observation should be randomly scattered and show no trends.

5.2.3 Constant variance

The variance of the residuals of an OLS model should be random, and unrelated to the value of the dependent variable or any of the independent variables. This condition is termed homoscedasticity. If this assumption is violated, then the data is said to be heteroscedastic.

Heteroscedasticity may be pure or impure (Frost, 2020a). Impure heteroscedasticity occurs in cases where an important factor has been left out of the model, and its unaccounted-for effect is absorbed by the error term. This would indicate a poor model, and should ideally be dealt with by identifying and including the missing variable(s). Pure heteroscedasticity occurs in cases where the model has been correctly specified, but the variance is naturally dependent on one of the variables. Pure heteroscedasticity is often present when one of the variables has a very large size range. For example, a model might predict the total pipeline length with a constant 10% accuracy, but a 10% error on 10 km of pipeline is much larger than a 10% error on 1 km of pipeline.

Heteroscedasticity can negatively impact a model in two ways. Firstly, the observations with larger residuals will have a greater influence on the least squares line. This could lead to biased

estimates of the regression coefficients, making the model less accurate for certain regions of the data. Secondly, the standard significance tests on the regression coefficients become unreliable. This may result in insignificant variables appearing to be significant. However, the reverse, that significant variables would appear insignificant, is unlikely (Frost, 2020a). The check for constant variance used in this study is described in Table 5-3.

Table 5-3: Check performed for constant variance assumption.

Check	Description
Residual Plots	Scatter plots of the residuals versus the dependent and each independent variable should be uniformly distributed and should show no thickening or narrowing. An example of a residual plot showing heteroscedasticity is displayed in Figure 5-3.

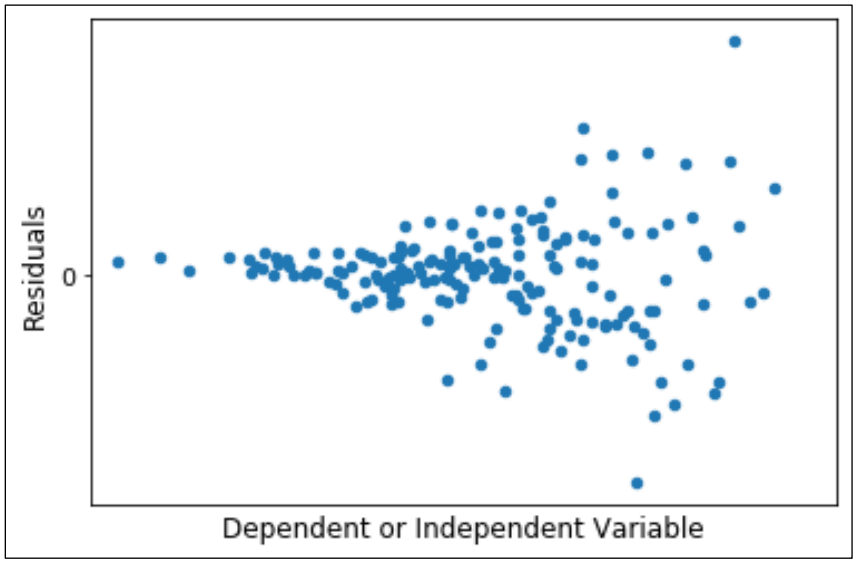


Figure 5-3: Residual plot displaying 'megaphone' shape of heteroscedasticity.

There are many possible solutions to heteroscedasticity. Two such solutions were investigated for this study. The first solution investigated was weighted least squares regression (WLS), which is a variation of OLS. In WLS, data points with higher residuals are down-weighted to reduce their disproportional impact on the regression coefficients. WLS regression is generally an appealing solution since it is simple to implement and does not affect the interpretation of the model. WLS regression is discussed in detail in Section 5.3. The second solution investigated was the

separation of the dataset into bins using the variable related to the heteroscedasticity, so that there are as many models as there are bins. This reduces the range of values in each bin, thus decreasing the effect of the heteroscedasticity in the models. Provided the sample size is large enough, this is a highly effective method because it is simple to implement and does not require any data manipulation.

5.2.4 Lack of multi-collinearity

No independent variables of an OLS model should be highly correlated, a condition known as multi-collinearity. Strong multi-collinearity leads to imprecise estimates of the regression coefficients, since the effect of each independent variable on the dependent variable cannot be isolated. Furthermore, if two variables are highly correlated, then they may be representing the same phenomenon, and it becomes redundant to include both in the model. For example, in a reservoir, the water volume and water level both indicate the volume of water in the reservoir, and multi-collinearity would be present if both were included in an OLS analysis. The two methods that were used to check for multi-collinearity in this study are described in Table 5-4.

Table 5-4: Checks for the multi-collinearity assumption.

Check	Description
Scatter Plots	Scatter plots of each independent variable versus each of the other independent variables should show that no two independent variables are strongly correlated.
Variance Inflation Factor (VIF)	VIF is a measure of the degree to which each independent variable is correlated with the other independent variables. A VIF of 1 indicates some multi-collinearity, which is normal; a VIF greater than 4 or 5 indicates unacceptable multi-collinearity (Montgomery & Runger, 2014).

If two or more variables exhibit multi-collinearity, only one of them should be used in the model. The one selected to remain should be the one with the strongest correlation with the dependent variable.

5.2.5 Normality

The last OLS assumption is that the residuals of an OLS model should be normally distributed. This assumption does not affect the ability of OLS regression to calculate the estimates of the regression coefficients. To this end, normality can frequently be neglected. However, the standard methods for testing regression significance and generating confidence intervals rely on the normality assumption. For these purposes, it is considered acceptable if the distribution is only close to normal, or 'normal enough'. The normality assumption becomes less important as the sample size grows, particularly for sample sizes larger than 30 or 40 (Montgomery & Runger, 2014). The normality assumption was important in this study since the sample sizes were frequently smaller than 40, and the significance tests on the regression coefficients were relied upon when determining which variables should be used in the models. The methods used in this study to check for normality are described in Table 5-5.

Table 5-5: Checks for the normality assumption.

Check	Description
Residuals Histogram	A histogram of the residuals should appear to follow a normal distribution.
Normal Probability Plot	A normal probability plot is a plot of the standardised residuals ('sample quantiles') versus a standard normal distribution with the same sample size ('theoretical quantiles') (NIST/SEMATECH, 2013). Figure 5-4 provides an example of a normal probability plot. A reasonably straight line on the normal probability plot indicates that the residuals are normally distributed.

Four additional indicators were used to support the official normality checks in Table 5-5, namely the skewness, kurtosis, Omnibus test, and Jarque-Bera test. The skewness is a measure of the distribution symmetry and is 0 for a normal distribution. The kurtosis is a measure of the distribution peak height and has a value of 3 for a normal distribution. The Omnibus and Jarque-Bera tests both use the skewness and kurtosis to estimate how close the distribution is to normal. For each of these methods, a value of 0 indicates a normal distribution, and the probability that the test value indicates a normal distribution should be 1. Since these indicators were only used as secondary checks, they are not included in the results.

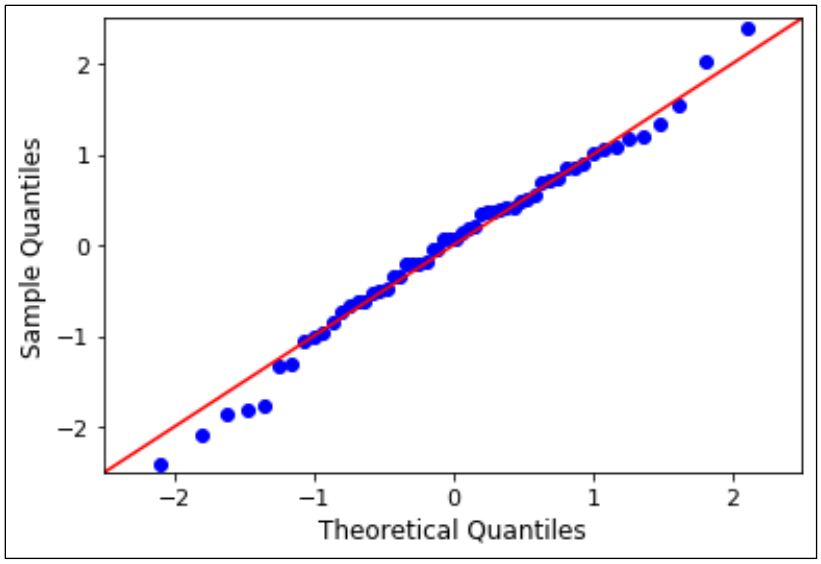


Figure 5-4: Example of a normal probability plot.

There are several reasons why residuals might be non-normal. One is that the linearity assumption was violated. If this is the case, then addressing the nonlinearity should solve the problem. Non-normality can also arise when the distributions of the dependent or independent variables are significantly non-normal themselves. In such cases, transformation of the offending variable to a normal distribution could solve this problem; however, any unnecessary data manipulation is normally not recommended. Another option is to not rely on the standard significance or confidence tests that assume a normal distribution. Instead, a more suitable distribution can be identified, and alternative statistical tests can be performed using this new assumption.

5.3 OLS Variation: Weighted Least Squares Regression (WLS)

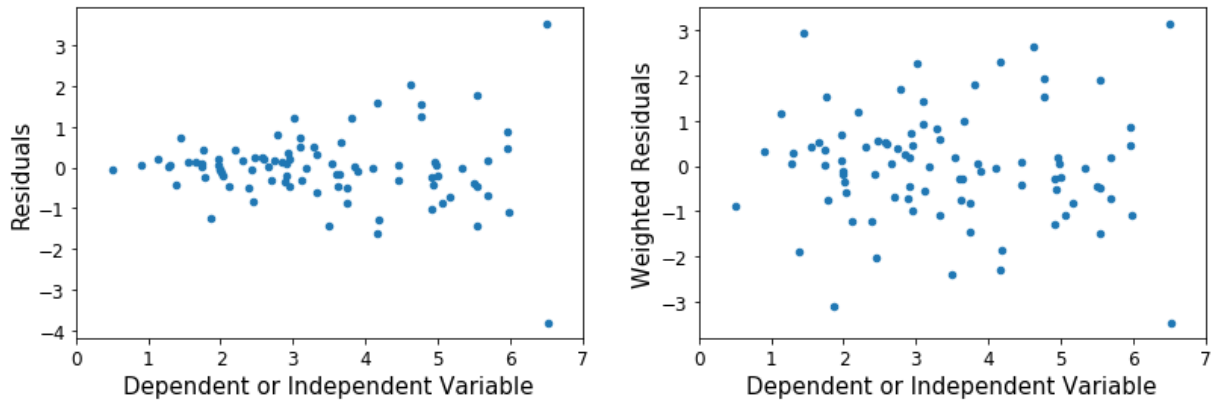
WLS regression was investigated in this study as a solution to the heteroscedasticity that was encountered in the data. WLS is a variation of OLS, in which data points with larger residuals are down-weighted to reduce their disproportional impact on the regression coefficients. As such, WLS allows for more accurate estimates of the regression coefficients. Pennsylvania State University (2018) provides a mathematical explanation of how WLS should be implemented. However, most statistical software packages provide WLS functionality in which only the dataset of independent and dependent variables, and the corresponding weightings, need be provided as inputs.

The challenge with WLS is that the perfect point weightings are not known, but must be estimated by the modeller. In this study, the variance of the residuals was related to the area size of the service zone. Accordingly, three common methods for estimating the weightings were used, listed in Table 5-6 (Pennsylvania State University, 2018). A limitation of WLS is that, when effective weightings cannot be identified, it is not a reliable solution.

Table 5-6: WLS weighting methods (Pennsylvania State University, 2018).

Weighting Method	Calculation
Weights 1	Weight by the inverse of the area.
Weights 2	Weight by the inverse of the variance. To estimate the variance, perform OLS as normal, and then regress the absolute values of the residuals against the *area*. The squares of the predicted values of the second regression are estimates of the variance.
Weights 3	Weight by the inverse of the variance. To estimate the variance, perform OLS as normal, and then regress the absolute values of the residuals against the *predicted values*. The squares of the predicted values of the second regression are estimates of the variance.

Since WLS is a variation of OLS, a WLS model must satisfy the same assumptions as an OLS model must, discussed in Sections 5.2.1 to 5.2.5. However, in the check for constant variance, heteroscedasticity will still be apparent in the residual plots since the variance of the actual residuals will still be non-uniform. Therefore, additional residual plots showing the *weighted* residuals versus the dependent and independent variables must be checked. The weighted residual plots indicate the variance of the residuals after the weighting has been applied. A uniform distribution would indicate that heteroscedasticity has been suitably addressed. The difference in the trends displayed by the residual and weighted residual plots is illustrated in Figure 5-5.



(i) Residual plot still indicates heteroscedasticity.

(ii) Weighted residual plot indicates heteroscedasticity has been addressed.

Figure 5-5: Residual and weighted residual plots after addressing heteroscedasticity.

5.4 Sample Size

In any statistical analysis, the sample size must be sufficiently large for the results to be considered reliable. There are many proposed methods for determining the required size of a data sample; the most common ones require desired level of error and the standard deviation of the data to be known (Montgomery & Runger, 2014). However, in multivariate regression analysis, when large datasets with multiple variables are considered, then such methods become complex and impractical. Therefore, it is quite common in multiple regression analyses that the sample size is simply checked against a rule of thumb; however, caution must be exercised when doing so since such rules of thumb may not be properly substantiated. For example, it is commonly said that at least five data points are required for each independent variable in a multiple regression analysis, but the origin of this recommendation is unclear.

However, there are also sample size rules of thumb that have been proposed in published research. Roscoe (1975) recommended that for multivariate regression, there should be at least 10 data points per independent variable, with the minimum sample size being 30 and the maximum sample size being 500. In a more recent study, Jenkins & Quintana-Ascencio (2020) attempted to determine the minimum sample size required to correctly match a model to a data shape. It was recommended that a minimum sample size of eight for data with very low variance, and a minimum sample size of 25 for data with a high variance, is required to accurately model the data.

For the sake of simplicity, in this study, a minimum sample size rule of thumb of 25 was adopted for multivariate regression models. Of course, this represented the minimum allowed number of data points, and the actual sample sizes were normally significantly larger than 25.

5.5 Outliers and Influential Points

There are a few reasons why a point might be considered an outlier. It could have an extreme value for one of the model variables. It could have a very large residual, indicating that it is far off the trend followed by the other points. Or, it might be considered an influential point, whose variable values and residual size might not be extreme, but both are large, so that the point has an unusually strong influence on the least squares line. There are no fixed rules about when or even if outliers should be removed, and it is normally left to the discretion of the modeller, considering the requirements of their specific study (Frost, 2020b). Therefore, a flexible approach to outlier removal was adopted for this study, and each outlier was judged on a case-by-case basis according to its influence on the relevant model.

Outliers were checked before compiling any model. Scatter plots of each independent variable versus the dependent variable were used to identify extreme values in any of the variables. Partial regression plots were used to identify influential points that might skew the regression coefficient for each independent variable. However, each data point was considered valuable, and outliers were only removed if it was clear that their absence would improve the quality of the model. Consequently, there were two conditions on which data points were removed. Firstly, a data point was removed if it had an extreme value for any of the variables, that made the model appear to be applicable to a much larger range of that variable than it really was. For example, for each land use, the largest area size was about double the second-largest area size, which made the area range appear twice as large. Secondly, in terms of large residuals, data points were not removed simply for having large residuals, as this would falsely make a model appear more accurate than it was. These points were only removed if they were also highly influential, and it was apparent that removing a single point would significantly change the slope of the least squares line.

A limitation of this approach was that it relied completely upon the discretion of the modeller. However, this approach was also tailored, so it potentially allowed some points to be retained which might have been removed if a formalised method was followed.

5.6 Model Building

When identifying which of a selection of candidate variables should form part of a model, the standard objective is to obtain the model with the highest prediction accuracy. However, a model with fewer variables is more efficient and easier to use. Model building is the process of obtaining the variable combination that best satisfies the two conditions of accuracy and simplicity. The most thorough approach is 'all possible regressions'. This entails compiling all the possible models with only one variable, two variables, and so on, up to a model with all the variables. Then, the model that best satisfies the objectives can be selected. However, this can be a time-consuming process.

A popular alternative to all possible regressions is stepwise regression. A variation of stepwise regression which was well suited to this study was backward elimination. In backward elimination, a preliminary model is compiled with all the candidate variables, and then insignificant variables are iteratively removed until only significant variables remain in the model. Before the backward elimination method for model building can be used, the desired significance level first has to be selected. The common default value is 95%, meaning that each variable must have a probability of being statistically significant of at least 0.95. The inverse is, therefore, that each variable must have a probability of being statistically *insignificant* (or p-value) of less than 0.05. For each regression coefficient, a hypothesis test is conducted where the null hypothesis is that the regression coefficient is equal to zero, which would indicate that the variable has no significance in the model. The p-value represents the probability that the null hypothesis is true, thus the probability that the variable is statistically insignificant. Therefore, a p-value of less than 0.05 indicates that a variable is acceptably significant in the model. Montgomery and Runger (2014) provides a thorough description of how the p-value is calculated.

To implement the backward elimination method, the preliminary model containing all candidate independent variables is first compiled, and then the associated p-value for each candidate variable is checked. If all the candidate variables have a p-value of less than 0.05, then the model remains as-is. However, if any variables have a p-value greater than 0.05, then the variable with the highest p-value above 0.05 is removed, and the model is re-compiled with the remaining candidate variables. This process is repeated until only variables with p-values of less than 0.05 remain.

In this study, a combined model-building approach was used. The backward elimination method was used as a first step. Then, depending on the remaining significant variables, models with additional variables removed were also compiled, in order to determine if an even simpler model could achieve similar results. Python software was used to calculate the p-values automatically. However, p-values become unreliable if the normality, constant variance, or lack of multi-collinearity assumptions of OLS are violated. Therefore, the p-values could only be relied upon if these assumptions were satisfied.

5.7 Model Building Variation: Principal Component Analysis (PCA)

Principal component analysis (PCA) is a technique in which the original independent variables are re-expressed as principal components, which are essentially new variables formed by weighted combinations of the original variables. The principal components are then used as the new independent variables in a standard OLS regression.

The concept behind PCA is that most of the variance in the independent variables is condensed into a few principal components (PCs). Fewer new independent variables can then be used in the OLS model. This reduces the dimension of the model, with minimal loss of meaningful information. Shlens (2014) provides a thorough mathematical explanation of how the PCs are created; but statistical software packages can generate the PCs with only the dataset of independent variables required as input. The output is the set of weightings used to convert the original independent variables into the PCs, such that the first PC accounts for most of the variation, and so forth. There are as many PCs as there are original independent variables. However, most of the variation in the dataset should be accounted for by the first few PCs. A common way to determine how many PCs should be used in a model is compiling the OLS model with only the first PC, determining the R^2, and iteratively adding components to the OLS model until each new PC does not significantly increase the R^2. Alternatively, the backward elimination method discussed in Section 5.6 can be used, where a model is compiled with all of the PCs, and then the insignificant PCs are iteratively removed until only the significant ones remain.

PCA is normally used when a dataset has a large number of possible independent variables, which makes the standard approach of variable selection inefficient. Another potential benefit of PCA exists for datasets where several of the independent variables are highly correlated. Rather than only retaining one of the multi-collinear independent variables and eliminating the rest, a new

variable can be determined consisting of a weighted contribution of the correlated variables. This could prevent some meaningful information from being lost, potentially resulting in a stronger model. Since there were nine correlated but distinct topography factors considered in this study, PCA was tested to see if this benefit could be realised. However, in regression analysis, a simpler model is always considered better. Therefore, the PCA model with more input factors would have to be significantly stronger than a model with only one topography factor for the additional data requirement to be justified.

5.8 Model Evaluation and Comparison Methods

There are numerous indicators available for evaluating and comparing the performance of regression models, each with its own strengths and limitations. This section discusses which evaluation and comparison indicators were used in this study, their significance, and their interpretation. Calculation of the indicators is not described here, since this is widely available information, and the indicators are easily generated by most statistical software packages.

5.8.1 Training and test datasets

Before compiling any model, the set of relevant data points was randomly split in an 80:20 ratio to form the training and test datasets, respectively. The training dataset was used to compile and evaluate the models, while the test dataset was reserved for evaluation purposes only. The test data was important for validating the model by proving that it worked for data that it had not been specifically fitted to. This provided a more objective measure of model performance.

5.8.2 Indicators for model evaluation and comparison

After compiling a model using the training dataset, three indicators were used for an intuitive evaluation of the model strength. These were the R^2, adjusted R^2, and test data R^2, which are described in Table 5-7. These indicators were useful for providing an easy-to-understand measure of how good a model was.

Table 5-7: Indicators used for intuitive evaluation.

Indicator	Interpretation
R^2	• R^2 is a goodness-of-fit metric describing the proportion of variance in the dependent variable explained by the model. • R^2 ranges from 0 to 1, where 0 indicates no correlation and 1 indicates perfect correlation. • Adding an independent variable always increases R^2, regardless of the model performance, therefore it is not reliable for model comparison.
Adjusted R^2	• The adjusted R^2 is R^2 with an adjustment factor applied, so that adding another independent variable will only increase it if the model is significantly improved. • A higher adjusted R^2 implies a stronger model.
Test Data R^2	• The test data R^2 refers to R^2 calculated for the test dataset. • It is used to check if the model performance is maintained for the test data. • The test data R^2 should be close to the training data R^2. It is acceptable if the test data R^2 is lower, but a large difference is a sign of overfitting.

While the adjusted R^2 in Table 5-7 is a good basis for model comparison, likelihood-based indicators provide an additional method for model comparison that might not necessarily identify the same model(s) as the best-performing. Therefore, the log-likelihood, AIC and BIC indicators, described in Table 5-8, were also used for model comparison, in order to verify agreement between them and the adjusted R^2. It is noted that these likelihood-based indicators have no specific meaning and imply nothing about how good a single model is; they can only be interpreted as relative values between models. Furthermore, these likelihood-based indicators are only applicable between models developed using the same sample points, and the same dependent variable.

Table 5-8: Indicators used for model comparison.

Indicator	Interpretation
Log-Likelihood	• The log-likelihood is an alternative goodness-of-fit metric to R^2. • A higher log-likelihood implies a better model. • Like R^2, it is biased towards models with more independent variables.
Akaike's Information Criterion (AIC)	• The AIC is a goodness-of-fit metric based on the same principal as log-likelihood, but with a penalty applied for more independent variables; thus, it balances model performance and complexity. • A lower AIC implies a better model, where a two-point difference is considered significant.
Bayesian Information Criterion (BIC)	• The BIC is similar to AIC, but with a heavier penalty applied for more independent variables. • A lower BIC implies a better model; where a two-point difference is considered significant.

If two models appeared equally good after assessing the indicators in Table 5-7 and Table 5-8, the simpler model was preferred. A model was considered simpler if it had fewer variables, or variables that were more easily obtainable. However, after a suitable 'best' model had been chosen, the indicators in Table 5-7 and Table 5-8 still did not provide much insight into how good the models actually were, or how their strength translated to accuracy of estimations. Therefore, the indicators presented in Table 5-9 were used to interpret the model accuracy.

The first two indicators in Table 5-9, namely the MAPE and 90% MAPE, expressed the average error size as a percentage and were thus easy to interpret and compare. The latter two indicators in Table 5-9, namely the MAE and RMSE, expressed the average error size as an absolute value, providing an important perspective on the results, but they were less useful for model performance comparison and assessment. Therefore the MAE and RMSE were relied upon to a lesser extent than the MAPE and 90% MAPE.

Table 5-9: Indicators used for interpreting model accuracy.

Indicator	Interpretation
Mean Absolute Percentage Error (MAPE)	• The MAPE is the mean of the absolute values of the errors expressed as percentages of the observed values. • The MAPE indicates the average error size, as a percentage of the dependent variable value.
90% Mean Absolute Percentage Error (90% MAPE)	• The 90% MAPE is the same as the MAPE, but it is calculated excluding the highest 10% of absolute percentage errors. • The 90% MAPE indicates how much the MAPE is skewed by the few largest errors.
Mean Absolute Error (MAE)	• The MAE is the mean of the absolute values of the errors. • The MAE indicates the average error size, in the units of the dependent variable. • The MAE is less sensitive to outliers than the similar indicator, root mean squared error (RMSE).
Root Mean Squared Error (RMSE)	• The RMSE is the square root of the average of the squared errors. • The RMSE indicates the average error size, in the units of the dependent variable.

5.9 Regression Methods Concluding Summary

In this chapter, the principles of regression analysis relevant to this study were described, along with how these techniques were applied in the context of this study. The following three chapters, Chapter 6, Chapter 7 and Chapter 8, are the Analysis chapters, which discuss the development processes of Study Outcome I, II and III respectively. In the Analysis chapters, it is assumed that the reader is *au fait* with the content discussed in Chapter 5 in terms of both the basic regression background knowledge and how the regression methods were adapted to this study. However, it is not required that the reader should have any knowledge of regression beyond what was discussed here.

Chapter 6
ANALYSIS FOR STUDY OUTCOME I: TOTAL PIPELINE LENGTH MODELS

From the data collection process detailed in Chapter 4, a dataset was collected for service zones of four major land use categories, namely: 'General Residential', 'Low Income Residential', 'Non-Residential', and 'Large'. The purpose of this segment of the analysis was to develop regression models for each of the four land use categories, expressing total pipeline length (dependent variable) as a function of the physical characteristics of a service zone (independent variables).

The model-development process consisted of the six major steps summarised in Figure 6-1. Steps 1 – 5 represented the coarse model development, in which major decisions were made regarding which variables, regression methods and solutions to problems should be used when developing the final models. Steps 1 – 5 therefore involved the compilation of numerous intermediate models, which were evaluated and compared to determine which elements should be included in the final models. Step 6 represented the model refinement, in which the outcomes from Steps 1 – 5 were applied to develop and refine the final models.

Regarding the software used for the analyses in this study, the basic data handling and inspection was done using Microsoft Excel. The regression analyses and all associated checks and visualisations were completed using the Python programming language, particularly the 'Statsmodels' and 'Scikit-learn' packages, which both had integrated regression functionality.

In the sub-sections that follow, each step in Figure 6-1 is discussed in detail, including the aim, approach, results and sub-conclusions of each step. Thereafter, the limitations of the model-development process are discussed, followed by a summary of the overall procedure.

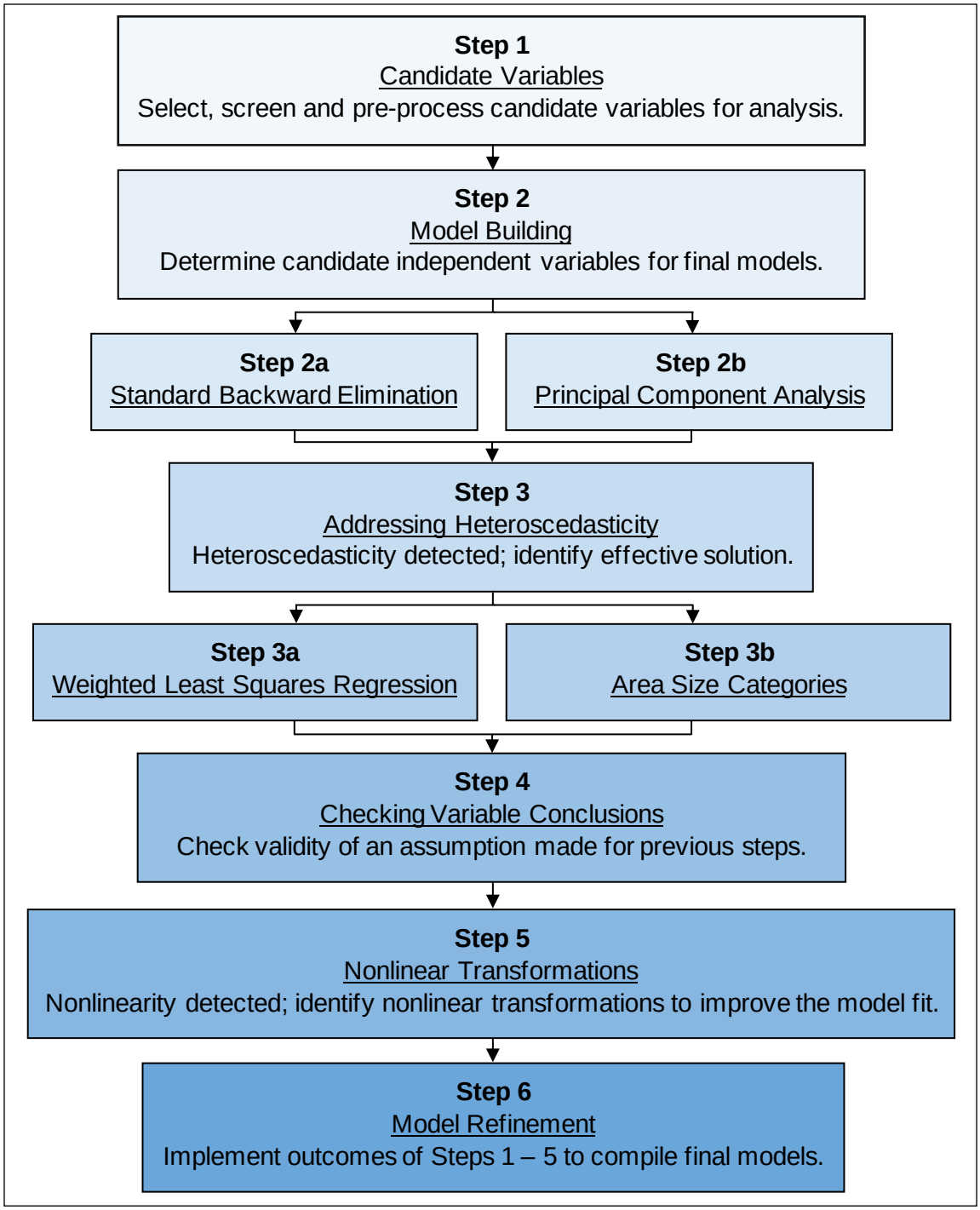


Figure 6-1: Study Outcome I development process.

6.1 Step 1: Candidate Variables

The first step was to screen the candidate variables to be used for developing the models. The total pipeline length was the desired dependent variable, and the set of 14 independent variables available for use consisted of the following: plane area, peak daily dry weather flow (PDDWF), number of unit hydrographs (UHs), circularity ratio, centroid-mouth relative radius, mean perimeter slope, mean basin slope, Melton's ruggedness number, surface area ratio, total relief, mean relief, elevation standard deviation, ruggedness number, and deviation from mean elevation. Table 4-3 in Chapter 4 provides definitions for all of the aforementioned variables.

The available variables were screened by inspecting the dataset, examining scatter plots, and performing preliminary regressions. Two problems were revealed at the onset that had to be addressed. Firstly, multi-collinearity (see Section 5.2.4) was present within two groups of variables. The first multi-collinearity group consisted of the variables plane area, PDDWF, and number of UHs. As noted in Section 5.2.4, the standard solution to multi-collinearity is retaining only the candidate variable with the highest influence on the dependent variable. Plane area had the highest individual correlation to total pipeline length and therefore remained as a candidate variable. To avoid the loss of potentially-valuable information, PDDWF and number of UHs were re-expressed by division by the plane area. This removed the correlation with the plane area, and they could be retained as candidate variables. However, the PDDWF per hectare and UHs per hectare still exhibited multi-collinearity, therefore they were grouped such that only one could be used in any given model. The second multi-collinearity group consisted of the nine topography factors. Since the topography factors all represented the same service zone characteristic, they were also grouped such that only one could be used in any given model. The best-performing variable in each of the two variable groups had to be established in the model-building process. The final candidate variables in their respective groups are presented in Table 6-1.

The second problem revealed by screening was the varying scales of the variables, which ranged in the order of 10^{1} to 10^{-3}. The varying scales caused the software to encounter difficulties when solving the regression coefficients. Variable scaling was therefore applied by pre-multiplying problematic variables by multiples of 10. The scale factors used are presented in Table 6-1. It is noted that variable scaling did not affect the model usage, but was incorporated in the regression coefficients.

Table 6-1: Summary of the candidate variables for the pipeline length estimation models.

Variable Group	Variables	Unit	Scale Factor
Y	Total pipeline length	km	-
X_1	Plane area	ha	-
X_2	PDDWF per hectare	kL/d/ha	-
	Unit hydrographs (UHs) per hectare	no./ha	-
X_3	Circularity ratio	m	x 10
X_4	Centroid-mouth relative radius	-	x 10
X_5	Mean perimeter slope	-	x 100
	Mean basin slope	-	x 100
	Melton's ruggedness	-	x 100
	Surface area ratio	-	x 10
	Total relief	m	-
	Mean relief	m	-
	Elevation standard deviation	m	-
	Ruggedness number	-	x 10 000
	Deviation from mean elevation	-	-

6.2 Step 2: Model Building

The second step was to determine which of the candidate independent variables from Table 6-1 should be used in the final models, a process normally referred to as model building (see Section 5.6). A two-pronged approach was followed. Firstly, the standard backward elimination method discussed in Section 5.6 was implemented to find the best-performing variable combination. Secondly, principal component analysis (PCA), as discussed in Section 5.7, was implemented to determine whether this alternative usage of the variables could yield better results than standard backward elimination. Each of the two methods required several models to be developed, and the results of the best models from each method were compared to decide which independent variables should be used in the final models. In the following three sub-sections, the backward elimination test, the PCA test, and the Step 2 sub-conclusions are discussed, respectively.

To simplify the analysis, the tests in Step 2 and Step 3 used only the 'General Residential' land use dataset. It was assumed that the same set of variables would also be the most significant for the other three land uses. This assumption was verified in Step 4. After four outliers were removed

from the original 'General Residential' sample of 240 data points, 236 data points remained, which were used in the Step 2 and Step 3 tests. The dataset was separated into a training set of 188 data points and a testing set of 48 data points.

6.2.1 Step 2a: Standard backward elimination

The aim of the backward elimination method was to determine which of the candidate independent variables were significant, and which significant variable in each variable group was the most significant. Thus, the best-performing combination of independent variables would be identified.

Using OLS regression, models were compiled for all 18 unique starting combinations containing one candidate variable from each variable group in Table 6-1, as well as for two additional starting combinations with fewer candidate variables. For each starting model, variables with $p > 0.05$ were removed by backward elimination, so that only the significant variables remained in the models. It is noted that all p-value conclusions were also verified using partial regression plots to visually assess the correlation strength, and to ensure that the conclusions were not influenced by outliers. Finally, the results of the end model versions comprising only significant variables were compared, in order to determine the best-performing combination of variables.

The full results are contained in Appendix E.1, including the starting combinations, end combinations, and corresponding performance results. The regression coefficients are omitted as these were not yet of interest at this stage. The most notable results from the backward elimination process are summarised in Table 6-2. Table 6-2 presents the variable combinations and key performance results of five models. The first four models (2a-E, 2a-F, 2a-N, and 2a-O) represent the four best-performing models from Step 2a. The fifth model, 2a-T, is included for comparison. The key performance results are presented in terms of the adjusted R^2, log-likelihood, AIC and BIC. As noted in Section 5.8, a stronger model is indicated by a higher adjusted R^2, more positive log-likelihood, and lower AIC and BIC.

Table 6-2: Summary of model results from Step 2a: Standard backward elimination.

Regression Model Number			2a-E	2a-F	2a-N	2a-O	2a-T
Variables	Y	Total pipeline length	x	x	x	x	x
	X_1	Plane area	x	x	x	x	x
	X_2	PDDWF per hectare	x	x			
		UHs per hectare			x	x	
	X_5	Total relief	x		x		
		Mean relief		x		x	
Results		Adjusted R^2	0.96	0.96	0.96	0.96	0.95
		Log-likelihood	-417	-410	-417	-411	-433
		AIC	841	828	842	829	870
		BIC	854	841	854	842	876

Firstly, model 2a-T, which used only plane area (variable group X_1) as an independent variable, had an adjusted R^2 of 0.95. This indicated that plane area was by far the most important independent variable, as it accounted for most of the variance in the dependent variable. However, the log-likelihood, AIC and BIC values showed that the models that incorporated additional independent variables performed better than 2a-T. This signified that while the plane area accounted for most of the variance in the dependent variable, the model estimates could be improved with information from additional independent variables.

The best four models from Step 2a (2a-E, 2a-F, 2a-N, and 2a-O) used additional independent variables from groups X_2 and X_5. To identify the best-performing variables from each group, the four models were compared. Concerning variable group X_2, models 2a-E and 2a-F, which used PDDWF per hectare, did not perform significantly better than their respective counterparts that used UHs per hectare, namely 2a-N and 2a-O. As noted in Section 5.8, the difference in log-likelihood, AIC or BIC must be at least two points to be considered substantial. Further tests were therefore required to conclude whether PDDWF per hectare or UHs per hectare was the better variable from group X_2. Concerning variable group X_5, however, models 2a-F and 2a-O, which used mean relief, did perform better than their respective counterparts that used total relief, namely 2a-E and 2a-N. Therefore, mean relief was identified as the best variable from group X_5, and selected to remain in the model. In summary, it was concluded from the backward elimination method that the best variable combination for the final models would be plane area, PDDWF per hectare *or* UHs per hectare, and mean relief (as in models 2a-F and 2a-O). The next step was to determine whether applying PCA would yield better results than models 2a-F and 2a-O.

It is noted that, during the backward elimination process, four variables were discarded from the dataset. The circularity ratio (variable group X_3), centroid-mouth relative radius (variable group X_4) and deviation from mean elevation (variable group X_5) had been removed with $p > 0.05$ from all or almost all of the models they were included in; therefore, these variables were regarded as totally insignificant. The ruggedness number (variable group X_5) required an additional test due to the nature of its definition, and was found to be an unreliable variable (see Appendix E.2). These four variables were given no consideration in the following steps.

6.2.2 Step 2b: Principal component analysis (PCA)

The aim of the PCA step was to determine whether, by allowing multiple independent variables per variable group to be used in a model, the results from the standard backward elimination could be improved upon. PCA was applied for the following three variable cases:

- All candidate variables bar the four removed in Step 2a,
- Only those of the above that were considered 'easily obtainable' in practice, and
- Only the best-performing four variables from Step 2a, including both PDDWF per hectare and UHs per hectare.

The latter two variable combinations were included to investigate whether simpler models that required less data could perform equally as well as the model that used all available variables. For all three variable cases, the principal components were generated, and OLS regression models were compiled using all the principal components as independent variables. Thereafter, for each of the three models, the principal components with $p > 0.05$ were removed by backward elimination until only significant principal components remained in the models.

The full results are contained in Appendix E.3, including the variable combinations, number of principal components, and all performance indicators. The regression coefficients and the make-up of the principal components are omitted as these were not of interest at this stage. The key performance results from the three PCA models are summarised in Table 6-3.

Table 6-3: Summary of model results from Step 2b: Principal component analysis.

Regression Model Number			2b-A	2b-B	2b-C
Variables in Principal Components	Y	Total pipeline length	x	x	x
	X_1	Plane area	x	x	x
	X_2	PDDWF per hectare	x	x	x
		UHs per hectare	x	x	x
	X_5	Mean perimeter slope	x	x	
		Mean basin slope	x		
		Melton's ruggedness	x	x	
		Surface area ratio	x		
		Total relief	x	x	
		Mean relief	x		x
		Elevation standard deviation	x		
Number of Significant Principal Components			7	3	3
Results		Adjusted R^2	0.96	0.96	0.96
		Log-likelihood	-406	-415	-409
		AIC	828	839	827
		BIC	854	852	840

Considering the three PCA models in Table 6-3, model 2b-A had the best log-likelihood value, but, according to the AIC and BIC values, model 2b-C achieved the best balance between accuracy and simplicity. Therefore, model 2b-C was considered the best PCA model. Model 2b-C was then compared to the best models from Step 2a, namely 2a-F and 2a-O. The adjusted R^2 values were equal, while the log-likelihood, AIC and BIC were marginally better for model 2b-C, but not by a value exceeding two points. Therefore, no evidence was found that PCA could improve on the results from Step 2a.

6.2.3 Sub-conclusions of Step 2

The most significant variables to be used for further model development were the plane area, PDDWF per hectare *or* UHs per hectare, and mean relief. The better variable between PDDWF per hectare and UHs per hectare still had to be established. The use of PCA did not lead to any model improvement and was not worth further consideration.

Additionally, the model-building process revealed that the OLS assumption of constant variance was violated, since the residual plots displayed the diagnostic 'megaphone' shape of

heteroscedasticity. It was a case of pure heteroscedasticity, where the size of the residuals increased with increasing plane area, as illustrated in Figure 6-2. As noted in Section 5.2.3, this would not indicate an incorrectly specified model, nor would it have been likely to have caused a significant variable to be ruled out as insignificant. Nonetheless, it did jeopardise the precision of the regression coefficient estimates, and it was addressed in the next model-development step.

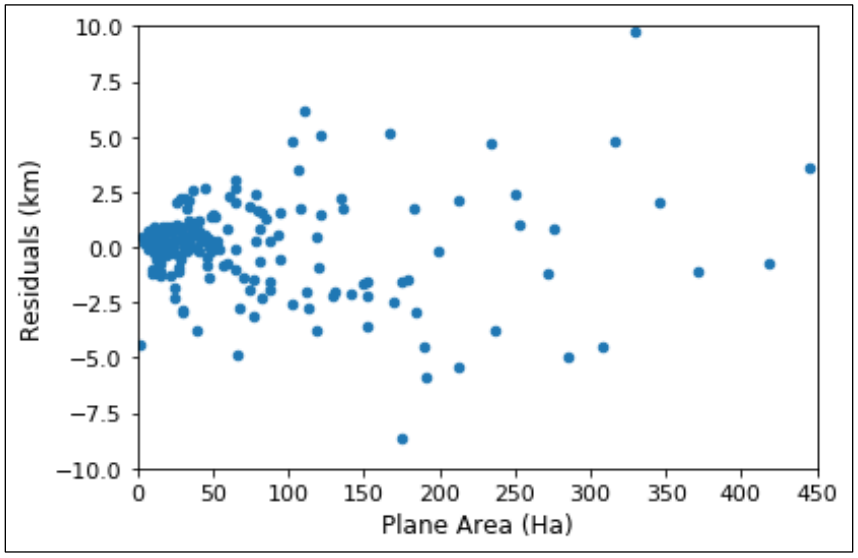


Figure 6-2: Plot of residuals versus plane area displaying heteroscedasticity related to plane area (from model 2a-O).

6.3 Step 3: Addressing Heteroscedasticity

The third model-development step was to reduce the impact of heteroscedasticity in the models. Two solutions were considered, namely weighted least squares regression (see Section 5.3) and area size categories (see Section 5.2.3). These are discussed in the following two subsections, followed by the Step 3 sub-conclusions.

6.3.1 Step 3a: Weighted least squares regression (WLS)

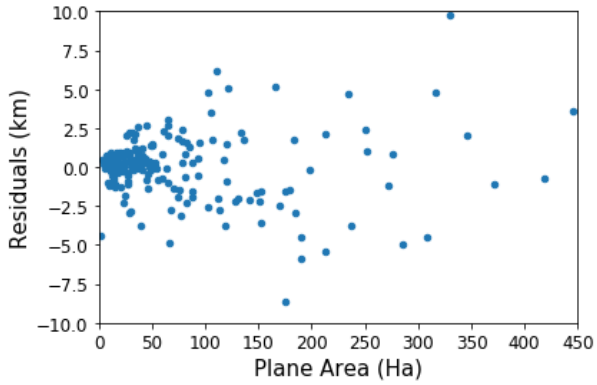
WLS is an appealing solution to heteroscedasticity, but only if effective weightings can be determined. A test was therefore performed to determine whether WLS could effectively address the heteroscedasticity, and which of the three weighting methods described in Section 5.3 (Weights 1, Weights 2, or Weights 3) was the most effective. An additional aim of this test was to determine which between PDDWF per hectare and UHs per hectare was the better variable.

Using WLS regression, a total of six models were compiled, using the three different weighting methods and the two possible combinations of the remaining independent variables (plane area, PDDWF per hectare *or* UHs per hectare, and mean relief). The full performance results are contained in Appendix E.4, but the key performance results are summarised in Table 6-4. It is noted that, unexpectedly, in regression model 3a-B, UHs per hectare had a p-value marginally greater than 0.05 and was removed from the model.

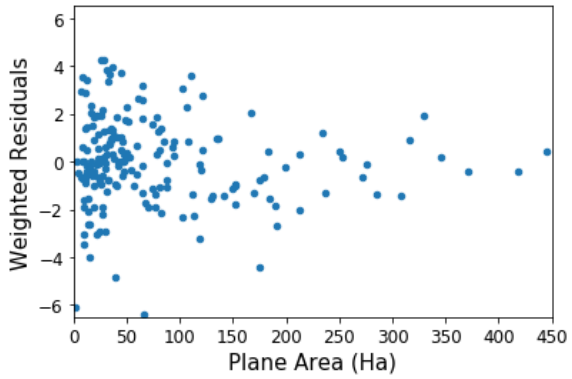
Table 6-4: Summary of model results from Step 3a: Weighted least squares regression.

Regression Model Number			**3a-A**	**3a-B**	**3a-C**	**3a-D**	**3a-E**	**3a-F**
Weighting Method			Weights 1		Weights 2		Weights 3	
Variables	Y	Total pipeline length	x	x	x	x	x	x
	X_1	Plane area	x	x	x	x	x	x
	X_2	PDDWF per hectare	x		x		x	
		UHs per hectare				x		x
	X_5	Mean relief	x	x	x	x	x	x
Results	Adjusted R^2		0.94	0.95	0.92	0.91	0.94	0.95
	Log-likelihood		-343	-333	-345	-350	-327	-321
	AIC		694	675	698	708	662	650
	BIC		707	688	711	721	675	663
	MAE test data (km)		1.27	1.20	1.44	1.42	1.33	1.27

To check whether the heteroscedasticity had been suitably addressed, the results from Table 6-4 were compared to the equivalent models from Step 2a in Table 6-2, namely models 2a-F and 2a-O. While the WLS models displayed a marginal decrease in the adjusted R^2 (from 0.96), the log-likelihood, AIC and BIC showed an average improvement of more than 100 points. It was concluded that WLS regression was an effective solution to the heteroscedasticity, and should be implemented when developing the final models. However, despite the improved performance, the residual plots revealed that all three weighting methods down-weighted the data points for large areas too severely, as illustrated in Figure 6-3. This introduced the risk that the models would be less accurate for large areas. Therefore, an additional method for addressing heteroscedasticity was investigated to minimise this effect, namely the introduction of area size categories. The area size categories method is discussed in Step 3b.



(i) Plot of total pipeline length residuals vs. plane area, displaying heteroscedasticity.



(ii) Plot of weighted total pipeline length residuals vs. plane area, displaying excessive down-weighting of large-area data points.

Figure 6-3: Residual plot and weighted residual plot for model 3a-F.

Regarding the weighting methods, the models that used Weights 3 performed the best overall in terms of the adjusted R^2, log-likelihood, AIC, and BIC. However, when considering an additional indicator, the MAE for the test data, Weights 1 showed a lower average error than Weights 3. Due to the uncertainty introduced by this, it was decided that the appropriate weighting method should be determined on a case-by-case basis when developing the final models in Step 6.

Regarding the better variable between PDDWF per hectare and UHs per hectare, models 3a-C and 3a-E, which used PDDWF per hectare, were respectively compared to models 3a-D and 3a-F, which used UHs per hectare (models 3a-A and 3a-B could not be included in the comparison since UHs per hectare was discarded from model 3a-B). However, while model 3a-C which used *PDDWF per hectare* performed better than model 3a-D which used *UHs per hectare*, model 3a-F which used *UHs per hectare* performed better than model 3a-E which used *PDDWF per hectare*. Therefore, the results were again inconclusive. Since PDDWF per hectare and UHs per hectare appeared to be equally strong variables, UHs per hectare was selected as the preferred variable on the basis of being simpler to calculate.

In summary, it was concluded from Step 3a that WLS should be implemented to address heteroscedasticity; that area size categories should be investigated; that the best weighting method should be determined on a case-by-case basis in the model refinement step; and that the variables to be used in the final models should be plane area, UHs per hectare and mean relief.

6.3.2 Step 3b: Area size categories

Based on the results of Step 3a, the introduction of area size categories was investigated to minimise the severe down-weighting of large areas. This involved separating the data points into categories according to the plane area, such that a separate model would be developed for each area size category. By reducing the area size range per model, it was expected that the weightings would be milder for the large-area data points.

In order to investigate the effectiveness of introducing area size categories, the 'General Residential' data points were separated into three area size categories, namely 0 – 40 ha, 40 – 100 ha, and 100 – 450 ha. The three area size categories contained 88, 50 and 48 data points respectively, after two additional outliers were removed. Using the winning independent variables from the previous steps, WLS regression models were compiled for the three area size categories. The weighting method was arbitrarily selected as Weights 3. The resulting models are summarised in Table 6-5. UHs per hectare again displayed a p-value marginally greater than 0.05 in model 3b-B, and was therefore removed from the model. This suggested that the significance of UHs per hectare could be related to the area size category.

Table 6-5: Summary of models from Step 3b: Area size categories.

Regression Model Number			**3b-A**	**3b-B**	**3b-C**
Weighting Method			Weights 3		
Area Size Range (ha)			0 – 40	40 - 100	100 - 450
Variables	Y	Total pipeline length	x	x	x
	X_1	Plane area	x	x	x
	X_2	UHs per hectare	x		x
	X_3	Mean relief	x	x	x

Table 6-5 does not provide means for model comparison, as none of the indicators were comparable in this case. The difference in the range of the total pipeline length for each area size category affected the relative scatter, thus biasing the R^2 indicators. The log-likelihood, AIC, and BIC are not comparable for models developed with different datasets, and the average error indicators also become difficult to interpret for models with different ranges of the dependent variable. Therefore, the only means of checking the effectiveness of implementing area size categories was inspecting the weighted residual plots for improved uniformity. The weighted

residual plots versus plane area for the Step 3b models are presented in Figure 6-4. The plots in Figure 6-4 displayed a more random and uniform scatter than the equivalent plot for the uncategorised data in Figure 6-3 (ii), indicating that area size categorisation was an effective solution. Therefore, it was concluded that area size categories should be implemented in the final models, with the actual category boundaries to be determined during the model refinement step.

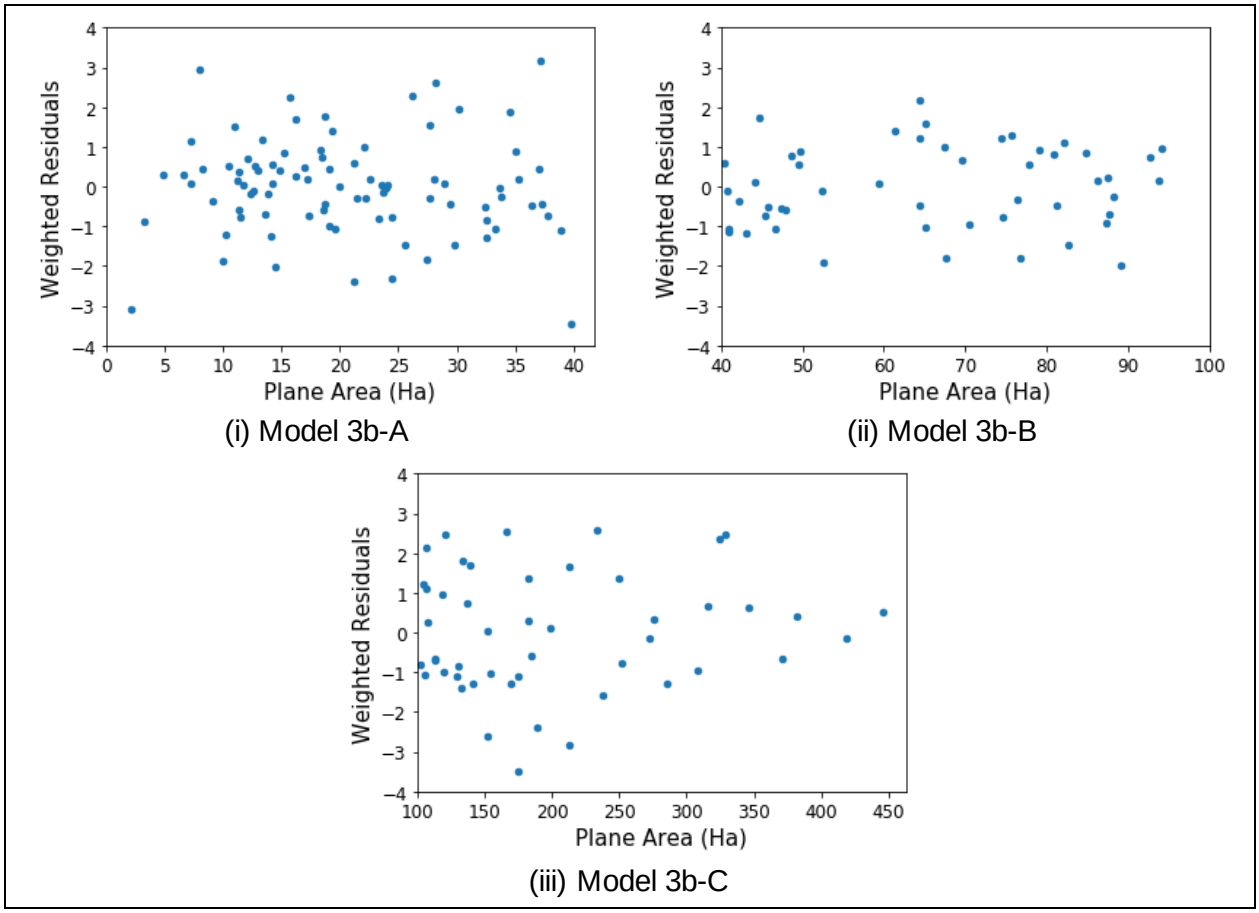


Figure 6-4: Plots of total pipeline length weighted residuals versus plane area for the different area size ranges in models 3b-B, 3b-C and 3b-D.

6.3.3 Sub-conclusions of Step 3

The best three variables to be used in the final models were the plane area, mean relief, and UHs per hectare. It was concluded that a combination of WLS and area size categories should be implemented to mitigate the effects of heteroscedasticity in the final models. The weighting method and the boundaries of the area size categories were left to be determined during the model refinement step.

6.4 Step 4: Checking Variable Conclusions

In Steps 2 and 3, it was assumed that best-performing candidate independent variables for the 'General Residential' land use would be the same for the other three land uses. The fourth step therefore involved a random verification as to whether this assumption was valid. 'Low Income Residential' was randomly selected as the verification land use. Using WLS regression with Weights 1 as the arbitrarily selected weighting method, models were compiled for 18 different starting combinations of candidate independent variables. The starting combinations were the same as those used in Step 2a, including the candidate independent variables which were discarded before Step 2b (except for the ruggedness number, which was not reliable). Thereafter, backward elimination was used as in Step 2a, so that only the significant variables remained in the models. The full results are contained in Appendix E.5, including the starting combinations, end combinations, and corresponding performance results. The results from the three best-performing models are summarised in Table 6-6.

Table 6-6: Summary of model results from Step 4: Checking variable conclusions.

Regression Model Number			**4-F**	**4-M**	**4-N**
Weighting Method			Weights 1		
Variables	Y	Total pipeline length	x	x	x
	X_1	Plane area	x	x	x
	X_2	PDDWF per hectare	x		
		UHs per hectare		x	x
	X_4	Centroid-mouth relative radius		x	
	X_5	Total relief		x	
		Mean relief	x		x
Results	Adjusted R^2		0.96	0.96	0.96
	Log-likelihood		-147	-147	-146
	AIC		303	304	300
	BIC		313	316	310

The best-performing model based on the log-likelihood, AIC, and BIC was model 4-N. Model 4-N comprised the same three independent variables selected as the best during Steps 2 and 3, namely plane area, UHs per hectare and mean relief. Based on this outcome, the assumption that the same independent variables could be used in the models for all four land uses was considered safe.

It is noted that model 4-M contained the centroid-mouth relative radius as a significant variable. While this variable was removed as completely insignificant in the 'General Residential' land use, it showed greater overall significance for the 'Low Income Residential' land use. Nonetheless, it was not part of the winning variable combination and thus did not form part of the final models.

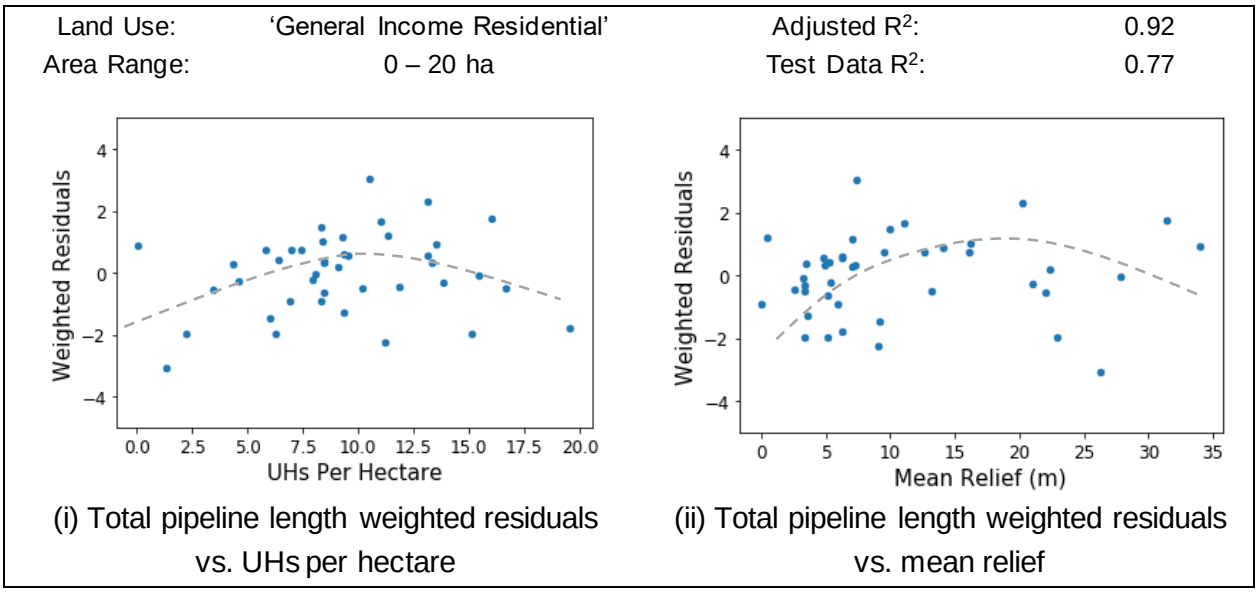
6.4.1 Sub-conclusions of Step 4

It was concluded that plane area, UHs per hectare and mean relief could safely be used as the best independent variables in the final models, for all four land uses considered.

6.5 Step 5: Nonlinear Transformations

With the major decisions regarding the model form having been made in Steps 1 – 4, the final models for the land use and area size categories could be developed. Preliminary area size categories were set for each land use, based on the number of data points available. The 'General Residential', 'Low Income Residential', 'Non-Residential' and 'Large' land uses could accommodate four, two, two and one area size category, respectively. This represented a total of nine model categories. WLS regression models were then constructed for the nine categories. However, when examining the partial regression and residual plots for each model, it was found that UHs per hectare and mean relief both exhibited a mild curved relationship with the total pipeline length. This phenomenon was not detected previously due to the high scatter of the data, and the obscuringly large number of data points used in Steps 1 – 4. Figure 6-5 provides an example of the curving weighted residual plots for one model ('General Income Residential', 0 – 20 ha); and Figure 6-6 provides an example of the curving partial regression plots for another model ('General Income Residential', 20 – 40 ha).

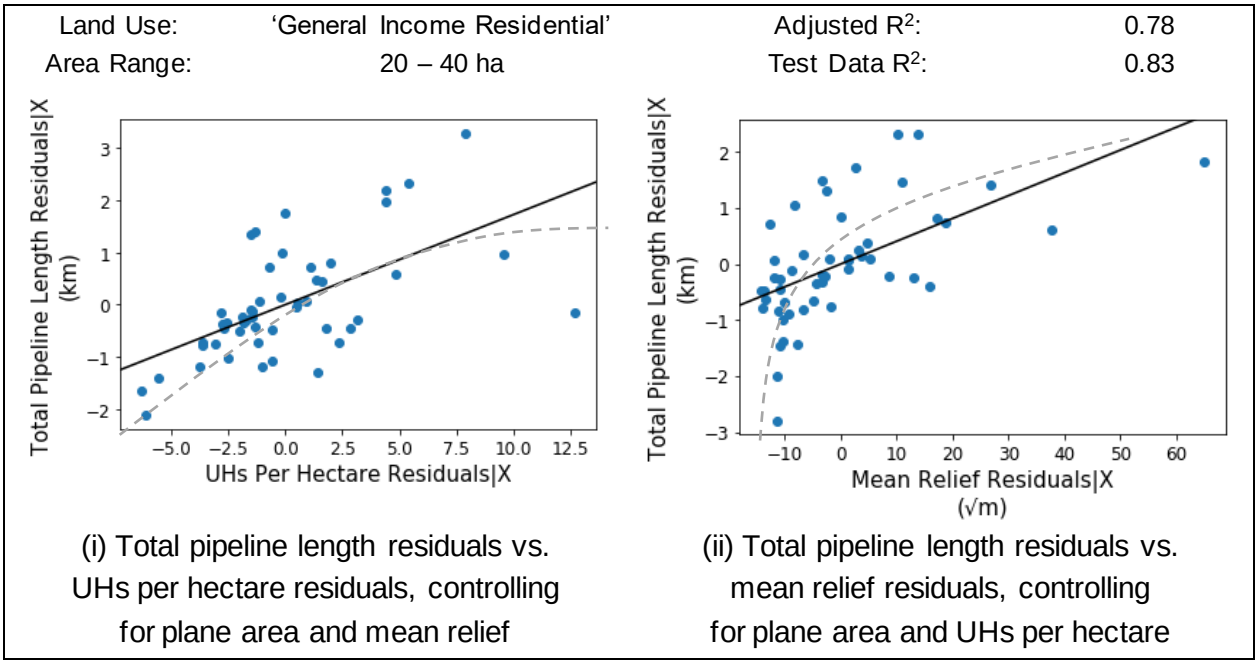
As discussed in Section 5.2.1, curvature in the residual plots and partial regression plots for an independent variable indicates a nonlinear relationship between that independent variable and the dependent variable, which must be addressed using an appropriate nonlinear transformation. Therefore, the ladder of re-expression method (see Section 5.2.1) was used to identify the nonlinear transformations for the UHs per hectare and the mean relief that resulted in the best visual data fit. Table 6-7 presents the transformations that were identified as the most suitable.



(i) Total pipeline length weighted residuals vs. UHs per hectare

(ii) Total pipeline length weighted residuals vs. mean relief

Figure 6-5: Residual plots displaying nonlinearity of UHs per hectare and mean relief.



(i) Total pipeline length residuals vs. UHs per hectare residuals, controlling for plane area and mean relief

(ii) Total pipeline length residuals vs. mean relief residuals, controlling for plane area and UHs per hectare

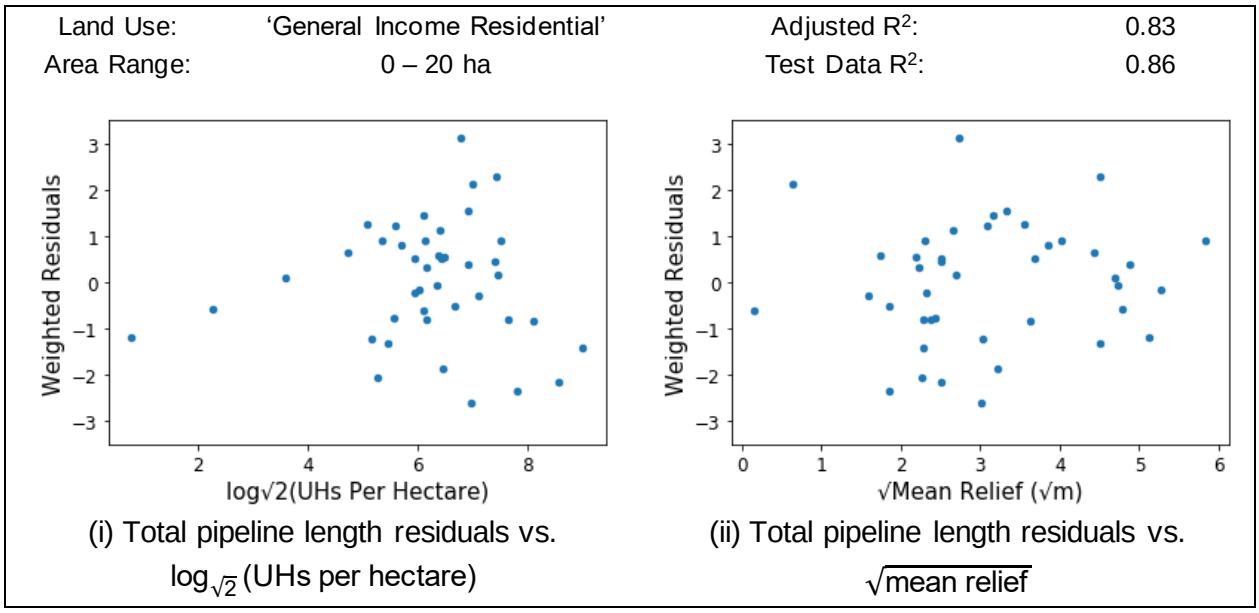
Figure 6-6: Partial regression plots displaying nonlinearity of UHs per hectare and mean relief.

Table 6-7: Nonlinear transformations applied to independent variables.

Independent Variable	Nonlinear Transformation
UHs per hectare	$\log_{\sqrt{2}}(x)$
Mean relief	$\sqrt{x}$

The nonlinear transformations improved the visual fit of the data for the different models significantly, which confirmed that this was a better representation of the true underlying relationship between the variables. Figure 6-7 and Figure 6-8 display the equivalent plots to Figure 6-5 and Figure 6-6 respectively, after the nonlinear transformations were applied to UHs per hectare and mean relief. The random scatter in Figure 6-7 and the relative straightness in Figure 6-8 indicate that the nonlinearity was satisfactorily addressed. Overall, between the nine models, there was little to no improvement in the performance indicator results. This unclear improvement is illustrated by the adjusted and test data R^2 values, which mildly decreased for the 'General Income Residential' 0 – 20 ha model, and mildly increased for the 'General Income Residential' 20 – 40 ha model. The test data R^2 remained high, which suggested that the nonlinear models were not overfitted to the data.

The newly-discovered nonlinearity of the underlying model form required the preceding variable significance conclusions from Steps 2 – 4 to be re-considered. The conclusions as to which variables were significant were still considered reliable, since the partial regression plots had been inspected for correlations before discarding any variables with $p > 0.05$. Furthermore, the models using plane area, UHs per hectare and mean relief still produced good results after the nonlinear transformations were applied. Therefore, these were still considered to be good selections of independent variables.

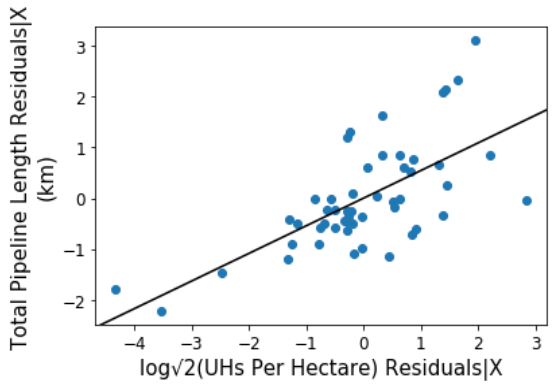


(i) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

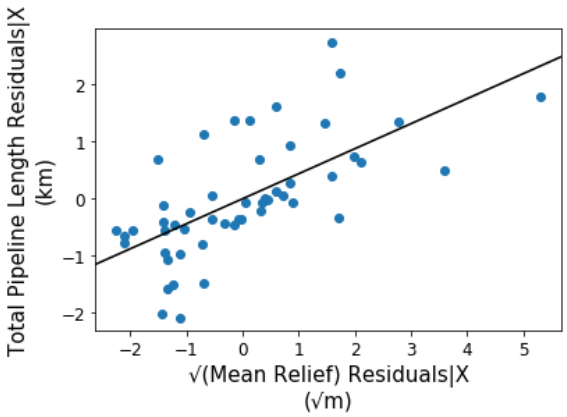
(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$

Figure 6-7: Example of residual plots displaying linearity after nonlinear transformations of the UHs per hectare and mean relief terms.

Land Use:	'General Income Residential'	Adjusted R^2:	0.79
Area Range:	20 – 40 ha	Test Data R^2:	0.91



(i) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare) residuals, controlling for plane area and $\sqrt{\text{mean relief}}$



(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area and $\log_{\sqrt{2}}$(UHs per hectare)

Figure 6-8: Example of partial regression plots displaying linearity after nonlinear transformations of the UHs per hectare and mean relief terms.

6.5.1 Sub-conclusions of Step 5

The nonlinear transformations of the variables mean relief and UHs per hectare (presented in Table 6-7) visually improved the fit of the models, and were considered to be better representations of the true variable relationships.

6.6 Step 6: Model Refinement

During the final step, the best model versions for the nine land use and area size combinations were developed. WLS regression was used with plane area, $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare) as the independent variables. The model refinement procedure illustrated in Figure 6-9 was followed iteratively for each model, in order to finalise the area size category boundaries, to ensure the correct outliers were removed, and to select the best weighting method for each case. The partial regression plots for the final models are contained in Appendix E.6, and summary sheets of the plots used to check the regression assumptions for each model are contained in Appendix E.7.

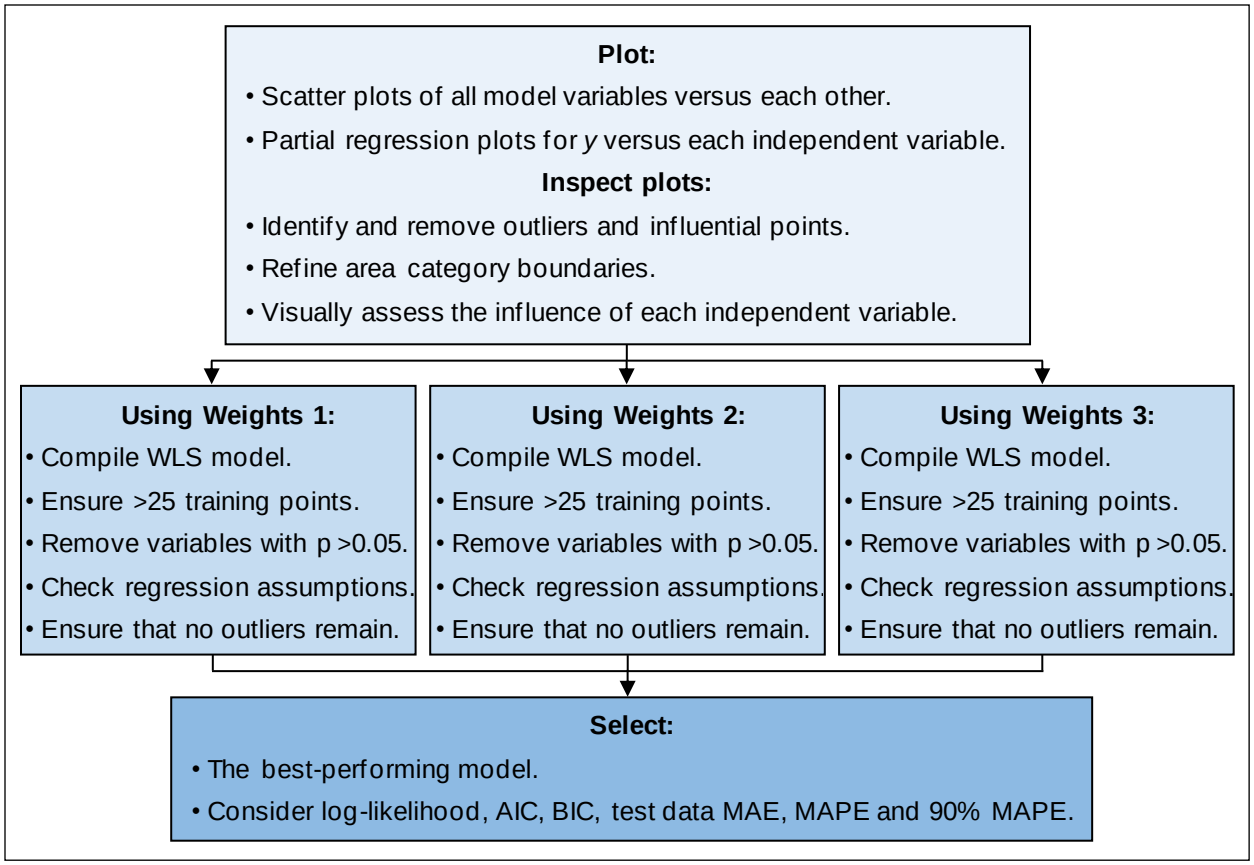


Figure 6-9: Model refinement procedure.

6.7 Additional Variable Availability Cases

It was observed that the plane area accounted for most of the variance in the total pipeline length, and the UHs per hectare and mean relief helped to improve the estimate. This suggested that reasonably accurate estimates would still be possible even in the absence of the latter two variables. Therefore, the same final models were re-compiled for the two additional cases of limited variable availability presented in Table 6-8.

Table 6-8: Variable availability cases.

Variable Case	Available Variables
A	Plane area; mean relief; UHs per hectare
B	Plane area; mean relief
C	Plane area

6.8 Limitations

A general limitation of multiple regression models is that developing these models is a subjective process, which requires the modeller to make certain decisions based on sometimes ambiguous results. The current study was no exception to this limitation, as several decisions were made based on the modeller's own interpretation of the results. It is therefore likely that a different modeller might have obtained final models that differed somewhat from the models developed here.

6.9 Analysis for Study Outcome I Concluding Summary

Developing the regression models for Study Outcome I was a multi-step process that involved selecting and grouping candidate variables, selecting significant variables using backward elimination, investigating PCA, addressing non-constant variance using WLS regression and area size categories, verifying assumptions, implementing nonlinear transformations, and lastly, compiling and refining the final models. The final models were created for the nine categories contained in Table 6-9, using WLS regression and three variable availability cases with plane area, $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare) as the independent variables. The final models and their performance results are presented and discussed in Chapter 9.

Table 6-9: Land use and area size categories of the final total pipeline length models.

Land Use Category	Area Size (ha)	Sample Size		
		Training Set	Testing Set	Total
General Residential	0 – 20	42	11	53
	20 – 40	40	11	51
	40 – 100	50	13	63
	100 – 450	48	13	61
Low Income Residential	0 – 40	53	14	67
	40 – 300	30	8	38
Non-Residential	0 – 40	36	9	45
	40 – 120	25	7	32
Large	0 – 160	20	5	25

Chapter 7
ANALYSIS FOR STUDY OUTCOME II: PIPELINE DIAMETER DISTRIBUTIONS

The purpose of this segment of the analysis was to develop pipeline diameter distributions that could be used to disaggregate the estimated total pipeline length into lengths of different diameters. The approach was simply to obtain the average percentage of total pipeline length per diameter within certain categories of similar networks. The development process required the categories to be set logically to obtain plausible diameter distributions. The approach followed is summarised in Figure 7-1. In the following sub-sections, each step in Figure 7-1 is described in detail, followed by the method limitations, and a summary of the overall process.

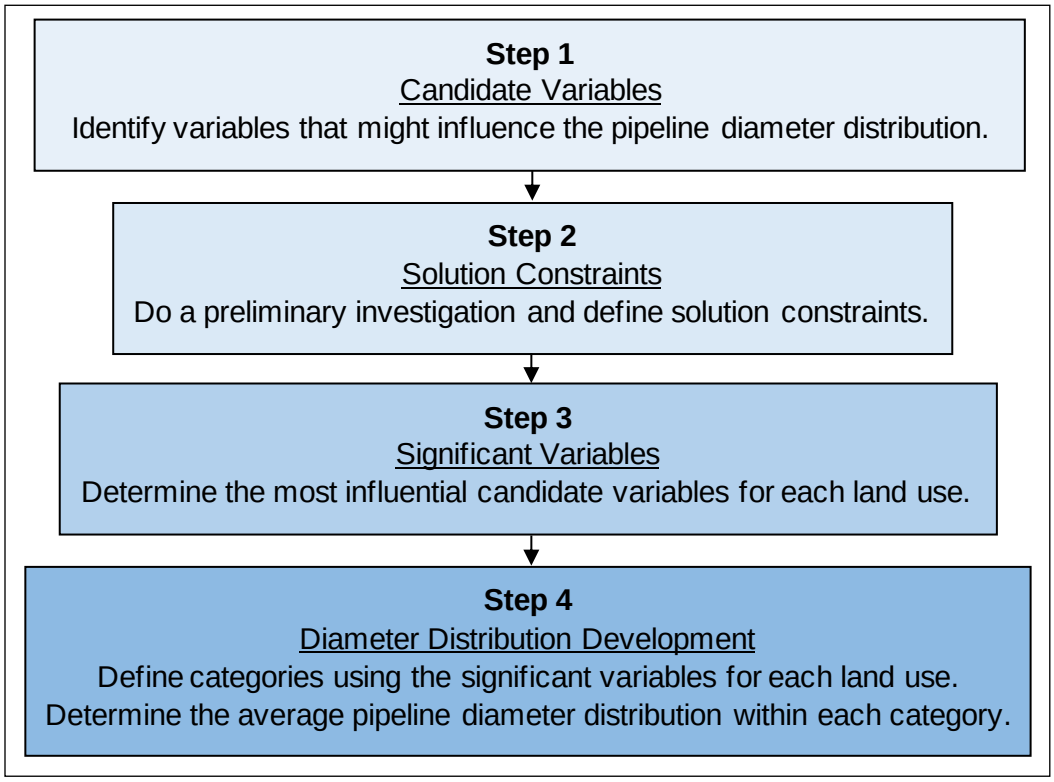


Figure 7-1: Study Outcome II development process.

7.1 Step 1: Candidate Variables

The desired outcome was the percentage of the total pipeline length per diameter size. Based on findings in the literature and the fundamentals of sewer network design, the factors that could potentially influence the diameter distribution were identified as the population size, area size, dwelling density, total design flow, per capita wastewater production, and topography. Considering the data available in this study, these factors could be accounted for by a combination of the following variables:

- Land use category
- UHs or UHs per hectare
- PDDWF or PDDWF per hectare
- Plane area
- Mean relief (the best topography factor from Chapter 6).

7.2 Step 2: Solution Constraints

A preliminary investigation was performed using trial categories and variable groupings in order to observe the effect on the resulting diameter distributions. The preliminary investigation revealed some constraints to the solution. Firstly, the diameter distributions of individual networks appeared to be fairly random, so that diameter distribution trends were only clear when there were many data points in each category. Therefore, the dataset was of insufficient size for subdivision according to all candidate variables from Step 1, such that there would be enough data points per category for logical trends to be revealed. Secondly, as in Study Outcome I, the PDDWF and number of UHs were so highly correlated to the plane area that it was impossible to isolate their effect from that of the plane area. Therefore, PDDWF *per hectare* and UHs *per hectare* were used. Again, these two variables were also so highly correlated that only one of them could be used for categorisation. And thirdly, the 'Large' land use dataset was too small, therefore it was combined with the 'Non-Residential' land use to form 'Non-Residential and Large'.

The solution constraints necessitated a formal test to be conducted to determine which of the candidate variables were the most influential, so that the limited possible number of categories could be set effectively. The test is described in Step 3.

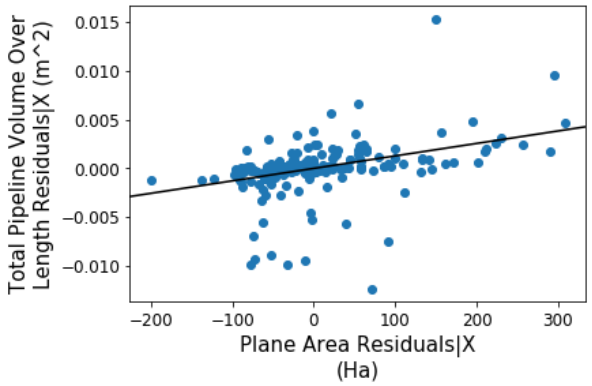
7.3 Step 3: Significant Variables

In order to determine the most influential candidate variables, the effect of each variable on the overall diameter distribution was evaluated by visual assessment of partial regression plots. The overall diameter distribution was represented using the total pipeline volume divided by the total pipeline length for each network. The total pipeline volume over length signified the average cross-sectional pipeline area of a network (indicating the average diameter). For each land use category, partial regression plots of the total pipeline volume over length versus plane area, UHs per hectare, PDDWF per hectare, and mean relief were plotted. The partial regression plots were analysed to assess which candidate variables had the strongest influence on the average cross-sectional pipeline area.

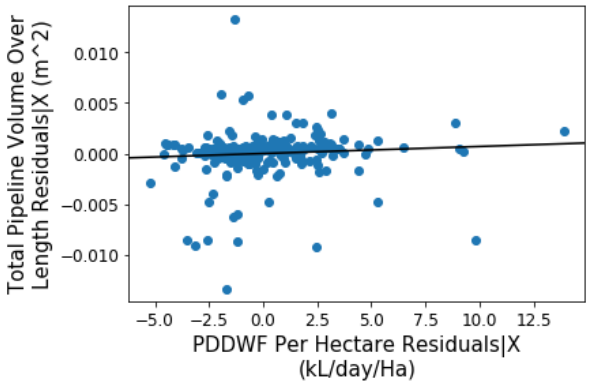
The partial regression plots for the 'General Residential' land use category are displayed in Figure 7-2. Contrary to expectation, the plots indicate that plane area was the only variable that exhibited any influence on the total pipeline volume over length for this land use.

The partial regression plots for the 'Low Income Residential' land use category are displayed in Figure 7-3. While PDDWF per hectare and mean relief showed a mild influence, plane area was again the only variable for which a substantial influence on the total pipeline volume over length could be seen.

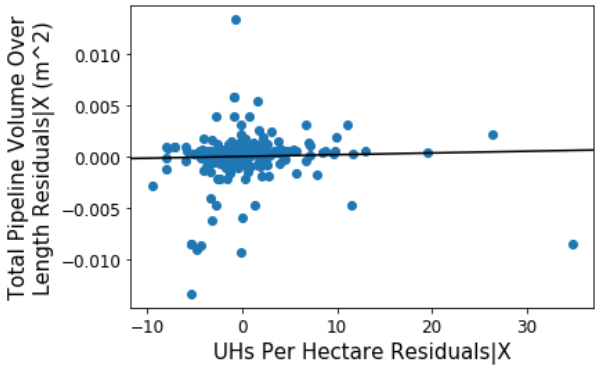
The partial regression plots for the 'Non-Residential and Large' land use category are displayed in Figure 7-4. While PDDWF per hectare did appear to have some influence, the trend was not considered consistent enough to be reliable. Therefore it was concluded that for this land use, plane area exhibited the strongest influence on the total pipeline volume over length, followed by mean relief.



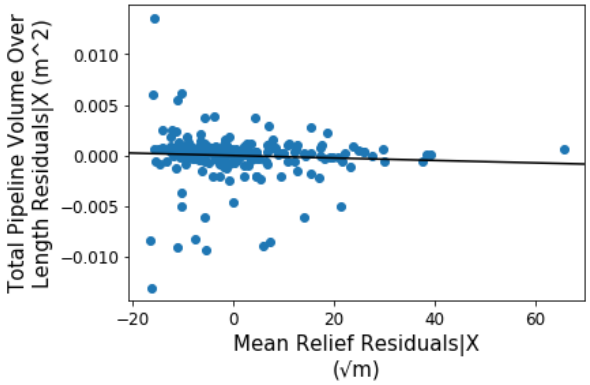
(i) Total pipeline volume over length residuals vs. plane area residuals, controlling for PDDWF per hectare[4] and mean relief



(ii) Total pipeline volume over length residuals vs. PDDWF per hectare residuals, controlling for plane area and mean relief



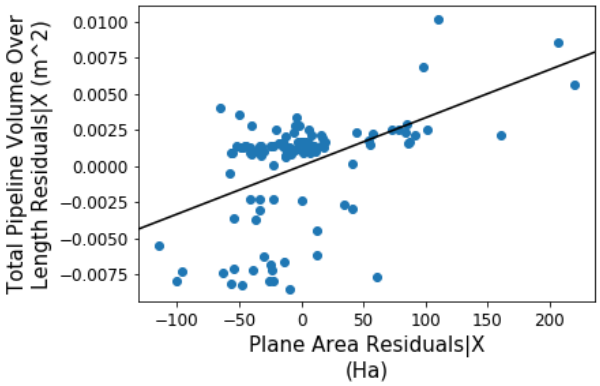
(iii) Total pipeline volume over length residuals vs. UHs per hectare residuals, controlling for plane area and mean relief



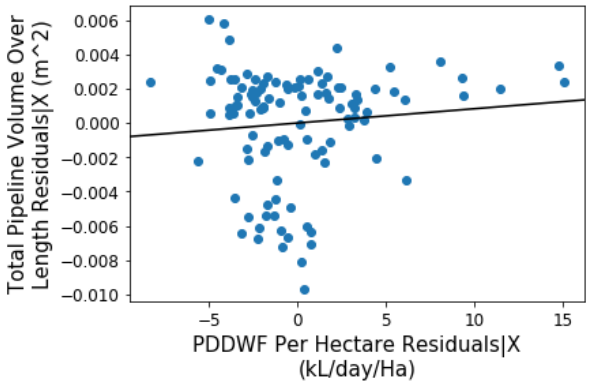
(iv) Total pipeline volume over length residuals vs. mean relief residuals, controlling for plane area and PDDWF per hectare

Figure 7-2: Partial regression plots showing the effect of the independent variables on the total pipeline volume over length ('General Residential' land use).

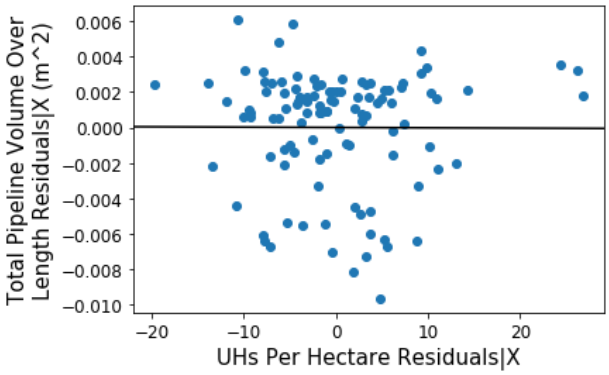
[4] Due to multi-collinearity, in partial regression plots showing the influence of plane area or mean relief, either PDDWF per hectare or UHs per hectare was controlled for, but not both. The plots controlling for UHs per hectare yielded the same conclusions as those controlling for PPDWF per hectare.



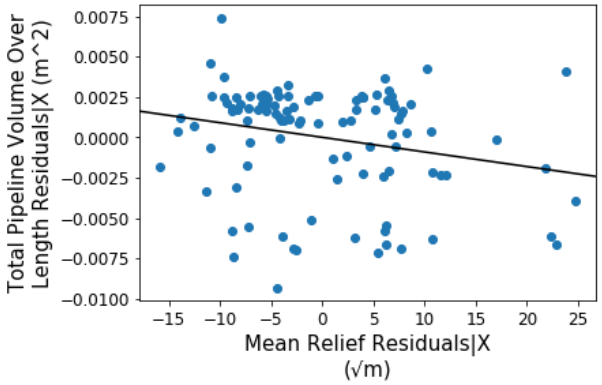
(i) Total pipeline volume over length residuals vs. plane area residuals, controlling for PDDWF per hectare and mean relief



(ii) Total pipeline volume over length residuals vs. PDDWF per hectare residuals, controlling for plane area and mean relief

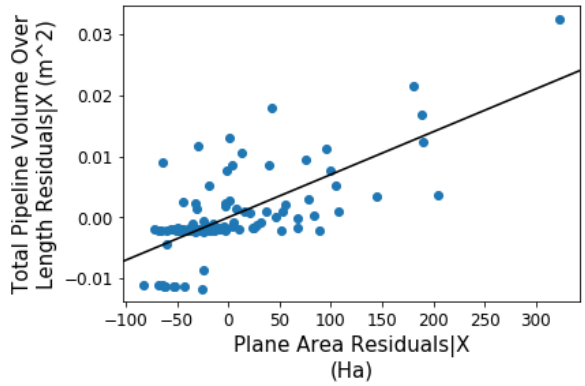


(iii) Total pipeline volume over length residuals vs. UHs per hectare residuals, controlling for plane area and mean relief

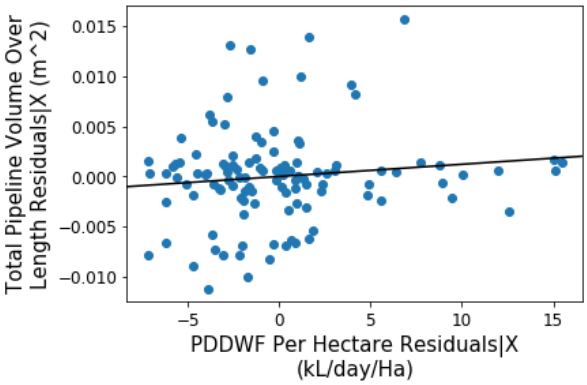


(iv) Total pipeline volume over length residuals vs. mean relief residuals, controlling for plane area and PDDWF per hectare

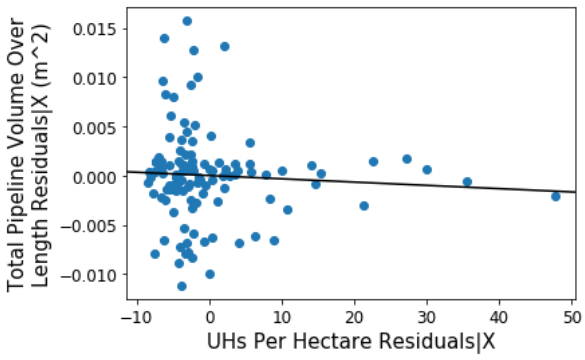
Figure 7-3: Partial regression plots showing the effect of the independent variables on the total pipeline volume over length ('Low Income Residential' land use).



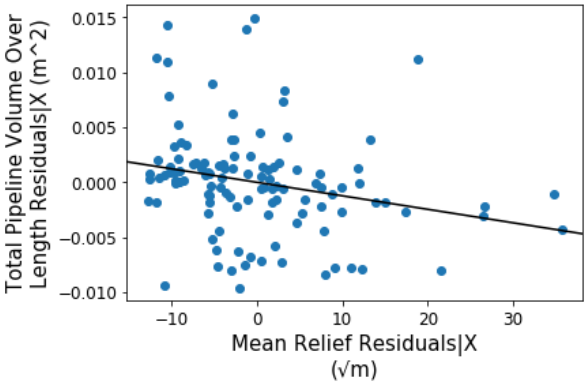
(i) Total pipeline volume over length residuals vs. plane area residuals, controlling for PDDWF per hectare and mean relief



(ii) Total pipeline volume over length residuals vs. PDDWF per hectare residuals, controlling for plane area and mean relief



(iii) Total pipeline volume over length residuals vs. UHs per hectare residuals, controlling for plane area and mean relief



(iv) Total pipeline volume over length residuals vs. mean relief residuals, controlling for plane area and PDDWF per hectare

Figure 7-4: Partial regression plots showing the effect of the independent variables on the total pipeline volume over length ('Non-Residential and Large' land use).

Based on the conclusions drawn from the partial regression plots, it was established that the 'General Residential' and 'Low Income Residential' land uses should be subdivided according to plane area only, while the 'Non-Residential and Large' land use should be subdivided according to plane area and mean relief. The plots were also used to help identify and remove any outliers, after which 454 sample points remained.

7.4 Step 4: Diameter Distribution Development

Finally, Step 4 entailed setting the number of categories and the boundaries thereof, based on the significant variables identified in Step 3 for each land use. For the 'General Residential' and 'Low Income Residential' land uses, the data points only had to be subdivided into area size categories. For the 'Non-Residential and Large' land use, the data points were first subdivided according to area size, and then according to mean relief. The expectation that guided the categorisation process was that larger and flatter areas should have a greater proportion of large-diameter pipes.

Scatter plots were used as visual aids when setting the category boundaries. The plots used showed the maximum pipe diameter versus plane area, and the total pipeline volume over length versus plane area. By indicating zones of relative homogeneity, these plots helped to ensure maximum distinction between bordering area size categories. Figure 7-5 and Figure 7-6 display the plots used for the 'General Residential' land use; the equivalent plots for the 'Low Income Residential' and 'Non-Residential and Large' land uses are contained in Appendix F.

Defining the appropriate categories was a trial-and-error process, guided by the following principles:

- There should be as many categories as possible, provided there are a reliable number of data points per category and the logical inter-category trends remain visible.
- There should be a noteworthy difference in the distributions between bordering categories, otherwise those categories should be combined.
- A category with many data points should only be further subdivided if this reveals a meaningful difference between the resulting sub-categories.
- The proportion of small pipes should decrease for larger and flatter areas.
- The maximum diameter should be greater for larger and flatter areas.

Overall, 17 categories in terms of land use, area size and topography were identified. The average diameter distributions were then generated for each of the 17 categories.

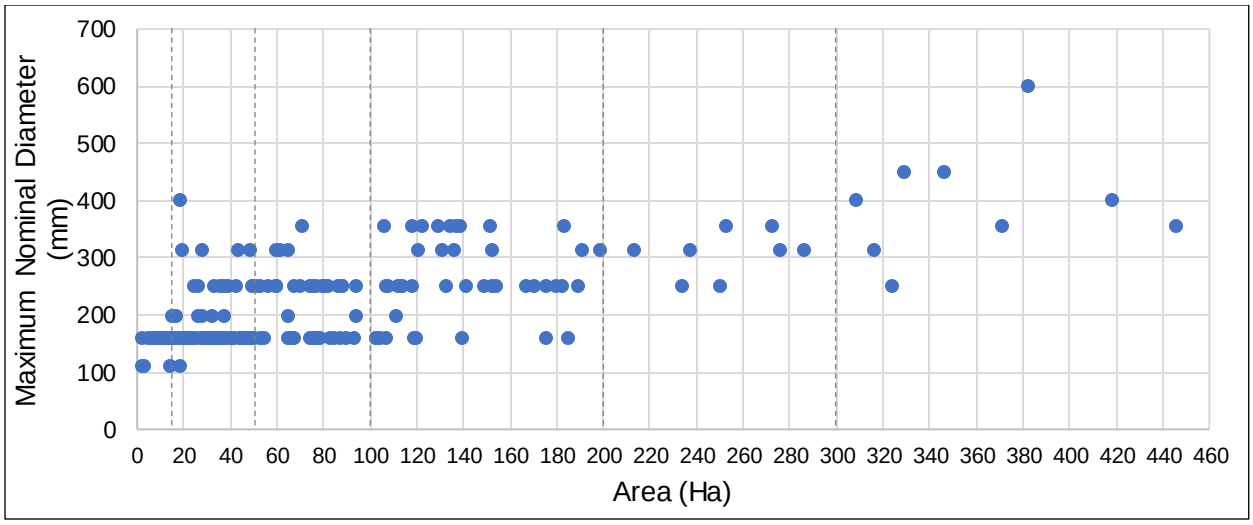


Figure 7-5: Maximum nominal diameter vs. plane area ('General Residential' areas).

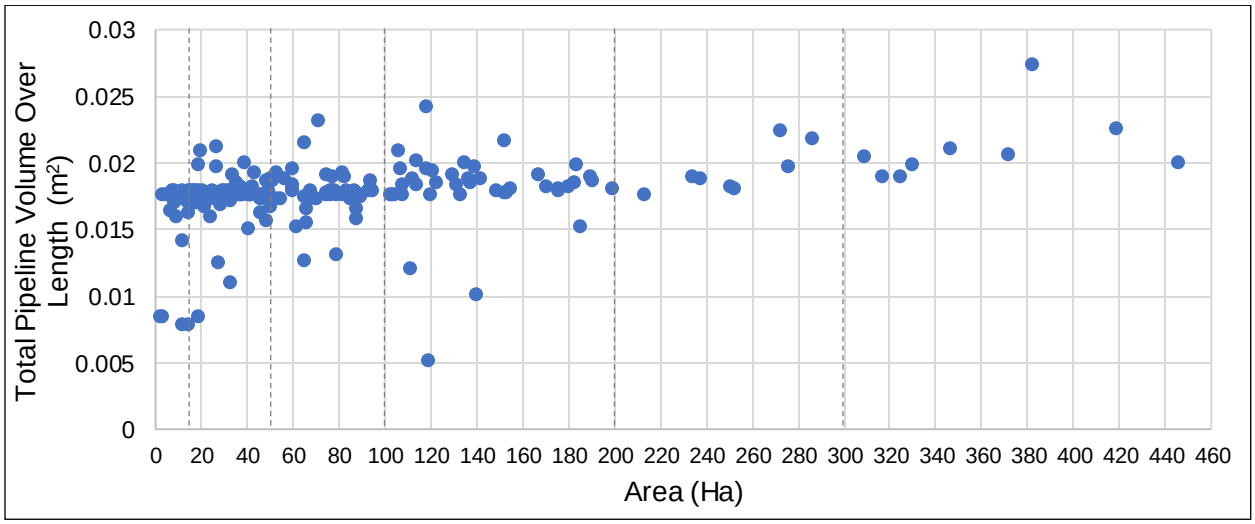


Figure 7-6: Total pipeline volume over length vs. plane area ('General Residential' areas).

7.5 Limitations

A limitation of this method was that it yielded somewhat subjective outcomes, although certain criteria and visual aids were used to determine category boundaries in a logical manner.

7.6 Analysis for Study Outcome II Concluding Summary

Developing the pipeline diameter distributions involved first identifying the potentially-influential variables, then identifying the most influential among them. The most influential variables were then used to develop suitable categories for which the average diameter distributions were determined. The final product was a set of pipeline diameter distributions for the 17 different land use, area size and topography categories contained in Table 7-1. The final distributions are presented and discussed in Section 9.2 of Chapter 9.

Table 7-1: Final categories for pipe diameter distributions.

Land Use Category	Area Size (ha)	Mean Relief (m)	Sample Size
General Residential	0 – 15	-	35
	15 – 50	-	95
	50 – 100	-	44
	100 – 200	-	42
	200 – 300	-	8
	300 – 450	-	9
Low Income Residential	0 – 20	-	45
	20 – 80	-	39
	80 – 150	-	14
	150 – 300	-	9
Non-Residential and Large	0 – 15	> 10	11
		≤ 10	16
	15 – 70	> 14	23
		≤ 14	32
	70 – 160	> 18	11
		≤ 18	10
	160 – 300	> 0	11

Chapter 8
ANALYSIS FOR STUDY OUTCOME III: MANHOLE DISTRIBUTION

The purpose of this segment of the analysis was to develop a method for estimating the number of manholes in a sewer network of a certain length. Originally, the objective was to obtain a generalised expression of the number of manholes per kilometre of pipeline. However, guidelines for manhole placement suggested that other factors may play a role in the manhole frequency. Therefore, an approach was taken that aimed to account for the influence of other variables on the manhole distribution in a network. The approach followed is summarised in Figure 8-1. In the following sub-sections, each step in Figure 8-1 is discussed in detail, followed by the method limitations, and a summary of the overall process.

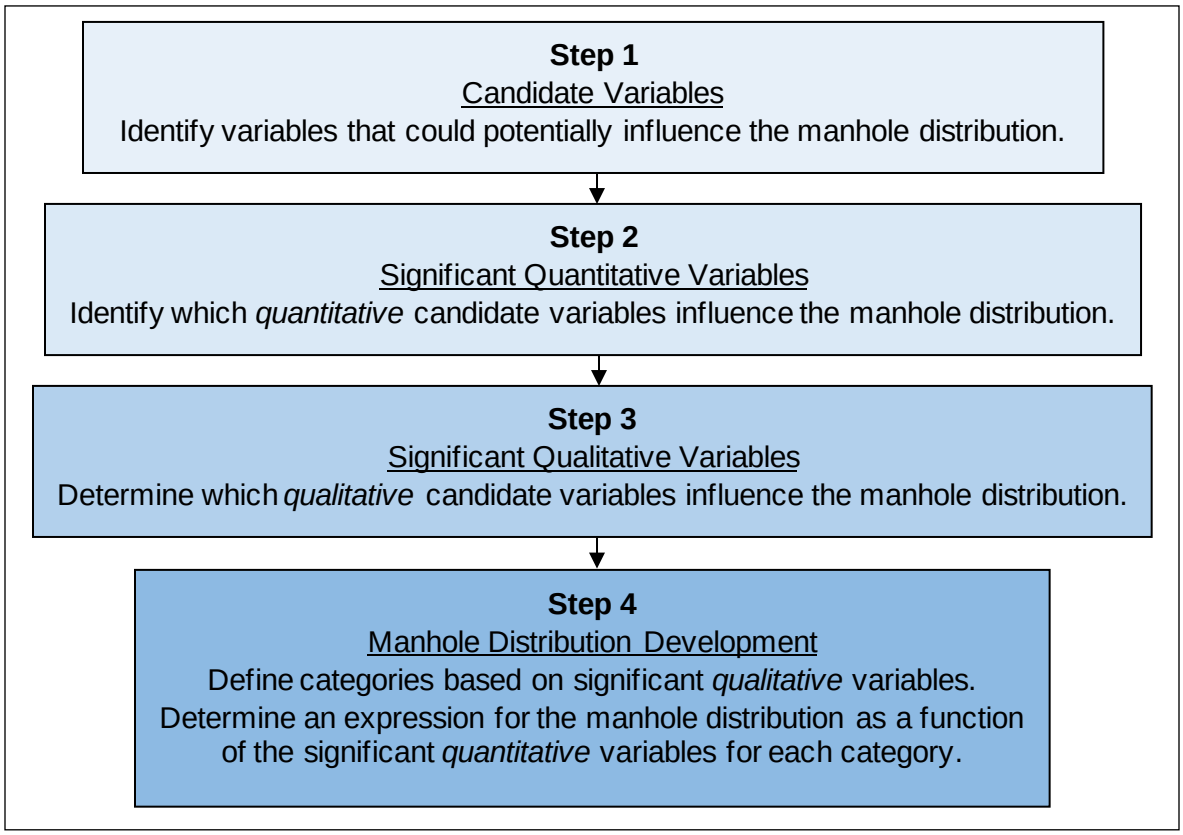


Figure 8-1: Study Outcome III development process.

8.1 Step 1: Candidate Variables

As noted in Section 2.2.3, according to DHS (2019), manholes must be placed at all pipeline junctions; gradient, diameter or direction changes; and at road crossings. Otherwise, the distance between manholes should not exceed 100 – 150 m. For steep sections, the manhole spacing is normally decreased, and for large-diameter sections it is normally increased. Based on these guidelines, many of the available variables were identified as potentially influential in the distribution of manholes. Therefore, all the candidate variables from the Study Outcome I analysis were also considered here. It is noted that the same variable groupings for preventing multi-collinearity still applied. The candidate variables for Study Outcome III are presented in Table 8-1.

Table 8-1: Summary of the candidate variables for the manhole distribution.

Variable Group	Variables	Unit	Scale Factor
Y	Number of manholes	-	-
X_1	Total pipeline length	km	-
X_2	Plane area	ha	-
X_3	PDDWF per hectare	kL/d/ha	-
	Unit hydrographs (UHs) per hectare	no./ha	-
X_4	Circularity ratio	m	x 10
X_5	Centroid-mouth relative radius	-	x 10
X_6	Mean perimeter slope	-	x 100
	Mean basin slope	-	x 100
	Melton's ruggedness	-	x 100
	Surface area ratio	-	x 10
	Total relief	m	-
	Mean relief	m	-
	Elevation standard deviation	m	-
	Deviation from mean elevation	-	-

8.2 Step 2: Significant Quantitative Variables

To account for the effects of multiple variables on the number of manholes in a network, an OLS model in the form of Equation 8-1 was used. This model expressed the number of manholes as a function of the total pipeline length and up to n other variables from Table 8-1. For each of the four land use categories, backward elimination was used to identify the significant variables.

$$Number\ of\ Manholes = \beta_0 + \beta_1 \times Total\ Pipeline\ Length + \beta_2 \times x_2 + \cdots + \beta_n \times x_n \qquad 8\text{-}1$$

For the independent variables in Table 8-1, there were 18 different possible combinations using one variable from each variable group. However, multi-collinearity was identified between plane area and total pipeline length. Since total pipeline length exhibited a stronger relationship with the number of manholes, plane area had to be discarded as an independent variable from the models. The models were compiled, and backward elimination was then performed for each starting combination, so that only the significant variables remained in the models.

Aside from total pipeline length, the only other variables that exhibited significance were the total relief, mean relief, and elevation standard deviation. The results of the models containing these independent variables were then compared. Table 8-2 shows the results for the 'General Residential' land use case; the results for all four land uses are contained in Appendix G. The results as to the best-performing variable combination in Table 8-2 were conflicting. Variable combination A was favoured by the adjusted R^2 and BIC; and variable combination C was favoured by the log-likelihood and AIC. The results for the other three land uses were similarly conflicting. Therefore, it was concluded that there was no evidence that including a topography factor (as in variable combinations B, C and D) could improve on the estimates made using only the total pipeline length (as in variable combination A). Therefore, aside from total pipeline length, no other quantitative variables from Table 8-1 were found to be significant in this investigation.

Table 8-2: 'General Residential' model results from Step 2: Significant quantitative variables.

Variable Combination		**A**	**B**	**C**	**D**
Independent Variables	Total pipeline length	x	x	x	x
	Total relief		x		
	Mean relief			x	
	Elevation standard deviation				x
Results	Adjusted R^2	0.98	0.96	0.96	0.96
	Log-likelihood	-971	-969	-968	-969
	AIC	1944	1944	1942	1944
	BIC	1947	1954	1951	1954

8.3 Step 3: Significant Qualitative Variables

Since the only significant quantitative variable was the total pipeline length, an OLS model of the form presented in Equation 8-2 was used to obtain the manhole distribution. The intercept term β_0 was forced to zero, so that the regression coefficient β simply represented the estimated number of manholes per kilometre of pipeline.

$$Number\ of\ Manholes = \beta \times Total\ Pipeline\ Length \qquad 8\text{-}2$$

However, the effects of land use and area size had not yet been investigated. Land use was a qualitative variable; and while area size could not be included as a quantitative variable in the previous step, it could still be represented qualitatively using area size categories. Therefore, using regression models of the form in Equation 8-2, the number of manholes per kilometre of pipeline was determined for different land use and area size categories.

From preliminary category groupings, two clear trends emerged. Firstly, the two residential land uses had similar manhole concentrations, as did the two non-residential land uses. And secondly, smaller areas displayed a higher manhole concentration than larger areas did. The latter trend was probably caused by larger areas incorporating more long, large-diameter pipeline sections, for which the distance between manholes generally increases, thus decreasing the average number of manholes per kilometre.

8.4 Step 4: Manhole Distribution Development

The final manhole distributions were calculated as the number of manholes per kilometre of pipeline, for different land use and area size categories. Based on the trends in Step 3, two land use categories were formed, namely ‘Residential’ and ‘Non-Residential’. Within each land use category, suitable boundaries for the area size categories were identified with the visual aid of scatter plots and partial regression plots. These plots were also used to identify and remove six outliers and influential points. Three area size categories were developed for each land use category, resulting in a total of six categories for the manhole concentrations. While there were sufficient data points for more categories to be formed, this would not have had much practical benefit, since the estimations between the existing categories were close. The final manhole distributions are presented in the Results chapter.

8.5 Limitations

A limitation of the manhole distribution development was that it only accounted for the total number of manholes, and did not attempt to quantify the prevalence of special structures that might occur at manholes, such as diversions, rodding eyes, top ends, or flow meters. Such special structures would in any case be minor features in the kind of upstream developments considered in this study. It is recalled from Table 4-3 that the total number of manholes for a sample network was defined as the number of manholes plus all other junction structures, since it was assumed that all such structures had an associated manhole.

8.6 Analysis for Study Outcome III Concluding Summary

Using backward elimination regression and data categorisation, it was found that the number of manholes could be expressed accurately as a function of the total pipeline length, land use and area size. Estimations of the number of manholes per kilometre of pipeline were developed using a simple linear regression model, for the six different land use and area size categories contained in Table 8-3. The final manhole distributions and associated performance results are presented and discussed in Section 9.3 of Chapter 9.

Table 8-3: Final categories for manhole distribution.

Land Use	Area Size (ha)	Sample Size		
		Training Set	Testing Set	Total
Residential	0-20	82	21	103
	20-50	82	21	103
	50-450	111	28	139
Non-Residential	0-30	41	11	52
	30-60	19	5	24
	60-160	22	6	28

Chapter 9
RESULTS

In this chapter, the results are presented in three major sections as follows:

- Study Outcome I: Results for Total Pipeline Length Models
- Study Outcome II: Results for Pipeline Diameter Distributions
- Study Outcome III: Results for Manhole Distributions.

In the three major sections, the final formulae or distributions for each Study Outcome are presented; the performance results thereof are evaluated and validated; and the findings are discussed in terms of overall performance and physical interpretation.

It is noted that the final dataset used to generate the Study Outcome components and corresponding results fell within the limits specified in Table 9-1. Therefore, the results presented and discussed in this chapter can only be considered applicable to service zones with characteristics falling within the specified limits.

Table 9-1: Ranges of the independent variables for model development and evaluation.

Land Use Category	Plane Area (ha)	Mean Relief (m)	UHs per Hectare
General Residential	0 – 450	0 – 82	1.3 – 22.7
Low Income Residential	0 – 300	0 – 53	4.9 – 48.7
Non-Residential	0 – 120	0 – 52	0.4 – 21.0
Large	0 – 160	-	-

9.1 Study Outcome I: Results for Total Pipeline Length Models

The models for the estimation of total pipeline length were compiled using only three independent variables, namely plane area, mean relief and UHs per hectare. Individual models were developed for nine different combinations of land use and area size categories. Models for these nine categories were developed considering three variable availability cases: Case A, in which all three independent variables were available; Case B, in which plane area and mean relief were

available; and Case C, in which only plane area was available. The result was a set of 27 models for the different land use, area size, and variable availability combinations. In the following subsections, the model formulae are presented, accompanied by an evaluation of their performance and an interpretation of the results.

Variable Case A represents the best-performing models, which are recommended for use, whereas Cases B and C provide alternative models that can be used when less information is available, with an associated reduction in accuracy. Therefore, only the results for Case A are considered in detail here, and the results for Cases B and C are referred to in the appendices.

Finally, it is important to note that when interpreting the total pipeline length output of the models, this output is constrained by the definition of a sample network used in the collection of the data points. In Section 4.2, a sample network endpoint was defined as the first point receiving the full combined flow of the network. Therefore, the total pipeline length models also represent the total pipeline length before this convergence point. By extension, this implies that the short length of pipeline which connects the network endpoint to the nearest collector sewer should be accounted for separately, on an application-specific basis.

9.1.1 Total pipeline length model formulae

The model form for variable Case A is presented in Equation 9-1. The variables y and x_i are defined in Table 9-2, and the regression coefficients β_i are provided in Table 9-3. For each regression coefficient, three values are provided. The 'Average' value provides the estimate of the true coefficient. Using the average coefficient values would provide the most likely total pipeline length for a service area. The 'Lower confidence limit' and 'Upper confidence limit' represent the boundaries of a 95% confidence interval on the coefficients. These are provided to allow the minimum or maximum total pipeline length that could reasonably be possible for a service area to be estimated. It is also noted that in some model instances, UHs per hectare, and sometimes also mean relief, were not significant. In such cases, the regression coefficients are zero. The equivalent model formulae for variable Cases B and C are contained in Appendix H.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{x_2} + \beta_3 \log_{\sqrt{2}}(x_3) \qquad 9\text{-}1$$

Table 9-2: Case A model variables.

Symbol	Variable	Unit	Calculation
y	Total pipeline length	km	-
x_1	Plane area	ha	-
x_2	Mean relief	m	Table 4-3 in Section 4.5.
x_3	UHs per hectare	number/ha	Table D-1 in Appendix D.2.

Table 9-3: Case A model regression coefficients.

Land Use Category	Area Size (ha)	β_0	β_1	β_2	β_3
General Residential	0 – 20	*-2.694*	*0.134*	*0.040*	*0.167*
		-1.845	**0.157**	**0.154**	**0.254**
		-0.996	*0.180*	*0.268*	*0.340*
	20 – 40	*-5.809*	*0.109*	*0.258*	*0.334*
		-4.189	**0.155**	**0.455**	**0.469**
		-2.569	*0.202*	*0.653*	*0.604*
	40 – 100	*-1.791*	*0.075*	*0.189*	*0.000*
		0.329	**0.102**	**0.530**	**0.000**
		2.448	*0.128*	*0.872*	*0.000*
	100 – 450	*-10.301*	*0.099*	*0.950*	*0.000*
		-6.214	**0.114**	**1.765**	**0.000**
		-2.128	*0.130*	*2.580*	*0.000*
Low Income Residential	0 – 40	*-4.180*	*0.169*	*0.112*	*0.172*
		-2.974	**0.187**	**0.244**	**0.297**
		-1.769	*0.205*	*0.376*	*0.422*
	40 – 300	*-27.043*	*0.134*	*0.144*	*0.949*
		-17.693	**0.153**	**0.884**	**1.962**
		-8.343	*0.171*	*1.624*	*2.974*
Non-Residential	0 – 40	*-0.845*	*0.064*	*0.009*	*0.069*
		-0.454	**0.083**	**0.142**	**0.114**
		-0.062	*0.102*	*0.274*	*0.160*
	40 – 120	*-2.974*	*0.034*	*0.522*	*0.000*
		-0.972	**0.060**	**0.885**	**0.000**
		1.029	*0.087*	*1.248*	*0.000*
Large	0 – 160	*0.635*	*0.029*	*0.000*	*0.000*
		0.961	**0.045**	**0.000**	**0.000**
		1.287	*0.062*	*0.000*	*0.000*
		Key:	*Lower confidence limit* **Average** *Upper confidence limit*		

9.1.2 Total pipeline length model performance evaluation

Model performance was evaluated using the 'Average' regression coefficient estimates. While several evaluation indicators were used in the model development, the final model performance was best described from two viewpoints, namely the R^2 and mean absolute percentage error (MAPE). For each model, the results generated using the training dataset were considered for the model evaluation, and these results were validated by checking the test dataset results. The test dataset comprised 20% of the data points, which were not used to compile the model.

9.1.2.1 R^2 for model correlation strength

Table 9-4 presents the R^2 values for the training and test datasets, for variable Case A. Equivalent results for Cases B and C are contained in Table I-1 in Appendix I. R^2 provides a useful and intuitive representation of the model strength. However, it must be analysed with some caution, as it can be affected by the resolution of the scatter plot (the range of values on the axes).

Overall, the R^2 values ranged from moderately good to very good, in the order of 0.6 to greater than 0.9, and these results were validated well by the test data results. However, the 'Large' land use case was an exception, where the test data R^2 of 0.04 indicated a very poor performance on the test set. On closer inspection of the 'Large' model, out of the five test points, there were two major under-predictions, but the other three were estimated reasonably accurately. It is therefore plausible that the test data results were skewed by two extreme-value points.

Table 9-4: R^2 for Case A models using training and test datasets.

Land Use Category	Area Size (ha)	R^2	
		Training Data	Test Data
General Residential	0 – 20	0.84	0.86
	20 – 40	0.80	0.91
	40 – 100	0.61	0.80
	100 – 450	0.87	0.94
Low Income Residential	0 – 40	0.91	0.93
	40 – 300	0.94	0.98
Non-Residential	0 – 40	0.81	0.60
	40 – 120	0.75	0.62
Large	0 – 160	0.64	0.04

Figure 9-1 provides a visual representation of the R^2 for the 'General Residential', 0 – 20 ha model, in the form of scatter plots of the predicted versus observed total pipeline length. Similar plots for all Case A models are provided in Appendix J. The strong 1:1 trends confirm the generally high R^2 values.

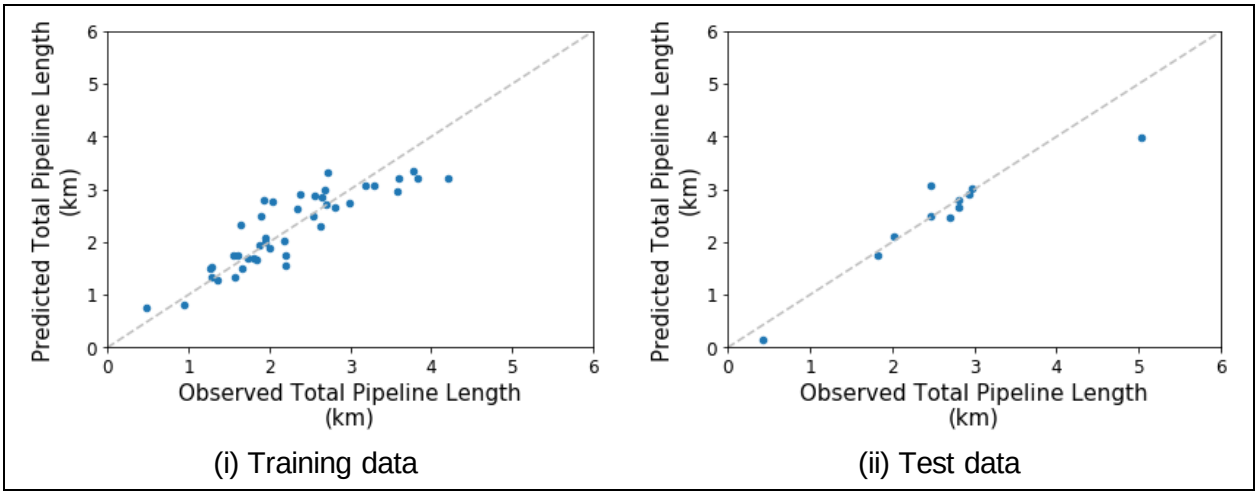


Figure 9-1: Predicted vs. observed total pipeline length ('General Residential', 0 – 20 ha).

9.1.2.2 Mean absolute percentage error (MAPE) for model accuracy

Table 9-5 presents the MAPE and 90% MAPE values for the training and test datasets, for variable Case A. The equivalent results for Cases B and C are contained in Table I-2 and Table I-3 in Appendix I. The MAPE indicates the average size of the absolute errors as percentages of the observed y-values. The so-called 90% MAPE was used to determine how accurate the models were in the best 90% of cases. The two indicators were used to obtain an intuitive indication of the model accuracy.

The MAPE values were generally good. Models for the 'General Residential' land use performed the best, with MAPE in the order of 10 – 15%. 'Low Income Residential' models also performed reasonably well, with a MAPE in the order of 10 – 20%. The MAPE for the 'Non-Residential' land use was overall a bit higher, in the order of 20 – 25%. The 'Large' category did not perform as well, with a MAPE in the order of 35%. On average, the 90% MAPE values were a few percentage points lower than the MAPE values, indicating that the model accuracy was considerably better in the 90% best cases. All results were validated by the test data, including the 'Large' land use.

Table 9-5: MAPE and 90% MAPE for Case A models using training and test datasets.

Land Use Category	Area Size (ha)	MAPE (%)		90% MAPE (%)	
		Training Data	Test Data	Training Data	Test Data
General Residential	0 – 20	14.8	12.6	11.0	5.3
	20 – 40	12.6	9.3	10.6	6.5
	40 – 100	13.9	13.1	11.1	11.9
	100 – 450	13.4	9.6	10.9	7.7
Low Income Residential	0 – 40	19.9	17.3	14.7	11.2
	40 – 300	10.2	7.7	8.3	7.0
Non-Residential	0 – 40	25.2	22.0	19.1	18.6
	40 – 120	18.9	20.8	15.1	15.8
Large	0 – 160	35.0	30.6	22.0	23.9

Overall, the performance results suggest that the model prediction accuracies range from good to moderate, with the 'General Residential' land use performing the best on average, and the 'Large' land use performing the worst. The generally poor performance of the 'Large' land use model is likely due to the combined effects of the small dataset (25 data points but only 20 training points), and the fact that the 'Large' sample networks sometimes contained partial sections of private industrial networks captured in the source models. This inconsistency could also explain why the area size was the only independent variable with a measurable influence for this land use.

9.1.3 Total pipeline length model physical interpretation

The final model form (Equation 9-1) contains nonlinear terms for mean relief and UHs per hectare. This model form indicates that the total pipeline length is expected to increase with increasing area size, mean relief, and UHs per hectare. More fundamentally, then, this means that the total pipeline length is expected to increase with increasing area size, relief, and density of contributing users, which is a logical conclusion. For the mean relief and UHs per hectare, an increase in either of these variables is associated with an increase in total pipeline length, but at a decreasing rate. In the case of UHs per hectare, this outcome could be physically interpreted as the required length of each new connection to a network becoming progressively shorter as a network changes from sparse to dense. For mean relief, such an intuitive interpretation is not clear, but the nonlinear relationship does seem reasonable.

Additionally, the final models show that UHs per hectare is not a significant variable in the larger area size categories (the regression coefficients are zero). This outcome makes sense, since small service zones might include a single development with a specific layout and population density, but large service zones incorporate more developments with a variety of population densities. Therefore, for larger service zones, UHs per hectare approaches an averaged value, thus losing its influence. This phenomenon is illustrated in Figure 9-2. Overall, in addition to strong performance results, the total pipeline length models are logical in their physical implications.

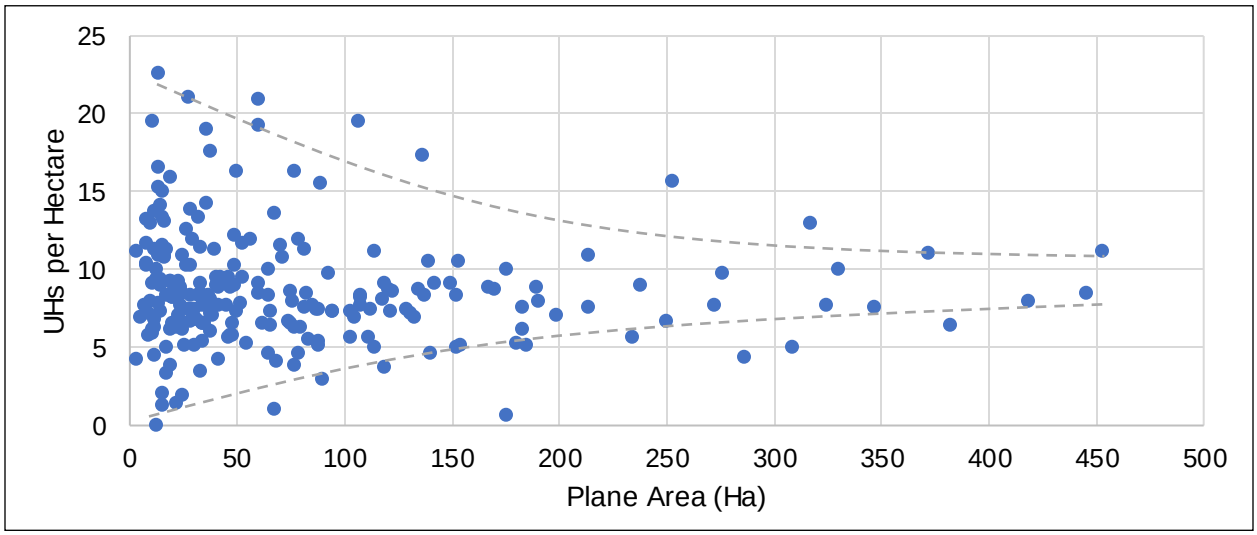


Figure 9-2: Stabilisation of UHs per hectare vs. plane area ('General Residential' areas).

9.1.4 Total pipeline length model summary

Overall, the Case A models do enable relatively accurate estimation of the total pipeline length. However, the estimation accuracy differs according to land use and area size category. The models and results from Cases B and C were not included in the main report, but interesting results arose. Some of the Case B and C models still performed very well despite the reduced number of independent variables, and therefore provide good alternatives in situations of reduced variable availability. Others, however, performed poorly and provide decidedly poor alternatives. For each of the 27 models developed for different land use, area size and variable availability combinations, the R^2 and MAPE results are provided. Therefore, whether a certain model is sufficiently accurate or not is dependent on the user's discretion considering the specific application. The given performance results should also guide the user as to whether a confidence interval estimate might be more suitable than the average expected pipeline length estimate.

9.2 Study Outcome II: Results for Pipeline Diameter Distributions

The diameter distributions were influenced by three variables, namely the land use, area size, and mean relief to a lesser degree. The distributions were calculated simply as the average percentage of total pipeline length per diameter, for different categories of the three influential variables. The result was a set of 17 unique pipeline diameter distributions for different categories of land use, area size and mean relief. The diameter distributions are presented, evaluated, and interpreted in the subsections to follow.

9.2.1 Pipeline diameter distribution tables

The diameter distributions for the 'General Residential', 'Low Income Residential', and 'Non-Residential and Large' land use categories are presented in Table 9-6, Table 9-7 and Table 9-8, respectively.

Table 9-6: Percentage total pipeline length per diameter ('General Residential' areas).

Area Size (ha)	Nominal Diameter (mm)											% Small Pipes*
	110	160	200	250	315	355	400	450	525	600	Total	
0 – 15	13.5	86.5									100	100
15 – 50	4.0	94.6	0.4	0.8	0.2						100	99
50 – 100	6.4	90.4	0.7	2.0	0.2	0.2					100	97
100 – 200	5.7	89.5	0.5	2.9	0.5	0.8					100	95
200 – 300	2.9	88.7	1.3	4.8	1.9	0.3					100	92
300 – 450	1.1	90.6	1.2	3.5	0.9	1.7	0.4	0.2	0.0	0.4	100	92

* Small pipes have diameter ≤ 160 mm.

Table 9-7: Percentage total pipeline length per diameter ('Low Income Residential' areas).

Area Size (ha)	Nominal Diameter (mm)												% Small Pipes*
	110	160	200	250	315	355	400	450	525	600	675	Total	
0 – 20	33.2	66.1	0.4	0.3								100	99
20 – 80	8.3	87.2	2.3	1.9	0.2							100	96
80 – 150	13.5	80.3	2.3	2.8	0.2	0.8						100	94
150 – 300	2.3	89.3	2.2	3.3	1.0	0.6	0.2	0.7	0.0	0.0	0.3	100	92

* Small pipes have diameter ≤ 160 mm.

Table 9-8: Percentage total pipeline length per diameter ('Non-Residential and Large' areas).

Area Size (ha)	Mean Relief (m)	Nominal Diameter (mm)											% Small Pipes *
		110	160	200	250	315	355	400	450	525	600	825	Total
0 – 15	> 10	45.5	54.5										100	100
	≤ 10	6.3	93.7										100	100
15 – 70	> 14	10.2	83.3	2.7	2.5	0.0	1.2						100	94
	≤ 14	5.2	84.5	1.3	5.7	2.0	1.2	0.1					100	90
70 – 160	> 18	2.1	89.9	3.1	4.4	0.5							100	92
	≤ 18	0.2	82.8	3.2	5.9	2.8	3.5	1.0	0.4	0.3			100	83
160 – 300	> 0	1.3	69.5	1.6	17.0	3.7	2.8	2.0	1.6	0.1	0.4	0.1	100	71

* Small pipes have diameter ≤ 160 mm.

9.2.2 Pipeline diameter distribution performance evaluation

There was no clear method for quantifying the pipeline diameter distribution performance meaningfully. However, insofar as logical distribution trends were concerned, the diameter distributions performed well. Overall, the proportion of small pipes decreased with increasing area size; the maximum nominal diameter increased with increasing area size; and in 'Non-Residential and Large' areas, flatter areas had a smaller proportion of small pipes and larger maximum diameters. In this sense the results were considered reflective of reality and thus fairly reliable.

Considering the reliability of the individual distributions, the most consistent trend was the percentage of small pipes (diameter ≤ 160 mm). The percentage of small pipes was always greater than 90% for the residential land uses, and greater than 70% for the non-residential land uses; and the individual values varied with area size and mean relief. This finding suggests that at least 90% (or 70%) of the total pipeline length can be expected, with a high level of confidence, to consist of pipes 160 mm in diameter or less. In effect, the majority of the pipeline network diameters could theoretically be estimated to within less than 100 millimetres of accuracy.

The distributions of the large pipes (diameter > 160 mm) were more random. This was possibly because the distribution of large pipes in a network is dependent on the specific network layout and the positions where the sub-networks converge. For example, a 450 mm converging with a 160 mm pipe might require a 525 mm pipe downstream of the convergence, but a 450 mm pipe

converging with a with a 315 mm pipe might require a 600 mm pipe downstream of the convergence, thus skipping the 525 mm category. Therefore the large-diameter distribution of any network is likely to deviate significantly from the average in most cases, which introduces considerable uncertainty for costing. However, based on the previous paragraph, large pipes account for less than 30% (largely less than 10%) of the total pipeline length, somewhat lessening the impact of the uncertainty. It is recommended that the distributions of the large pipes be used as a guide, but that they remain open to interpretation by the user based on the required level of conservativeness. To this end, the plots of maximum nominal diameter versus plane area for each land use contained in Appendix F, which were used to set area size category boundaries, may also be helpful in identifying the range of possible maximum diameters for an area.

9.2.3 Pipeline diameter distribution physical interpretation

The realistic trends displayed by the diameter distributions in relation to area size and mean relief were already discussed in Section 9.2.2. The distributions also varied according to the three land use categories. The most notable difference was in the small pipes. The 'General Residential', 'Low Income Residential' and 'Non-Residential and Large' land uses had, in that order, the highest to the lowest average proportion of small pipes. This trend was possibly due to the latter two having higher dwelling or flow production densities overall, as well as a possible compensation in low income areas for the more frequent occurrence of foreign matter in the sewer system. The 'General Residential' land use also generally had the fewest 110 mm pipes, which may have been influenced by municipal specifications. That being said, the most recent South African guidelines (DHS, 2019) stipulate a minimum diameter of 100 mm for reticulation pipes, and 200 mm for pipes in central business district (CBD) areas to allow for future densification. The new guidelines then suggest that certain non-residential service zones would have no small pipes at all.

9.2.4 Pipeline diameter distribution summary

The distributions did display the expected trends, which suggested that they do represent the *average* case for each category fairly well, but individual cases may deviate from this substantially. Therefore, the diameter distributions should be interpreted with caution. The most useful outcome was the proportion of small pipes, which allows upwards of 70% (or 90%) of a network's pipeline diameters to be estimated with relative certainty. It was recommended that for the remainder of the network, the large-diameter distribution results be interpreted conservatively.

9.3 Study Outcome III: Results for Manhole Distributions

The frequency of manholes and other junction structures (collectively referred to as manholes) along the pipeline length was influenced by two variables, namely land use and area size. Therefore, the manhole distribution was calculated as the number of manholes per kilometre of pipeline, for two land use categories with three area size sub-categories each. The manhole distributions are presented, evaluated and interpreted in the subsections to follow.

9.3.1 Manhole distribution table

The frequency of manholes and junction structures along the pipeline length are presented in Table 9-9. The 'Residential' land use category represents a combination of the 'General Residential' and 'Low Income Residential' land use categories from Study Outcomes I and II, and the 'Non-Residential' represents a combination of the 'Non-Residential' and 'Large' categories. The 'Average' value indicates the most likely true manhole frequency. The 'Lower confidence limit' and 'Upper confidence limit' provide the bounds of a 95% confidence interval on the manhole frequency, to allow the minimum and maximum number of manholes that could reasonably be expected to be determined.

Table 9-9: Distribution of manholes and other junction structures.

Land Use Category	Area Size (ha)	Number of Manholes per Kilometre of Pipeline		
		Average	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Residential	0 - 20	22.6	21.6	23.5
	20 - 50	21.3	20.4	22.1
	50 - 450	20.0	19.5	20.6
Non-Residential	0 - 30	20.5	19.1	22.0
	30 - 60	18.2	16.9	19.5
	60 - 160	17.0	15.8	18.1

9.3.2 Manhole distribution performance evaluation

Similarly to the total pipeline length models, the prediction performance of the manhole distributions was evaluated with the R^2 and MAPE for training dataset, and these results were validated using a test dataset.

9.3.2.1 R^2 for manhole distribution correlation strength

The R^2 for the predicted versus observed total number of manholes in each sample network are presented in Table 9-10. The high R^2 values, which all exceed 0.95, suggest that the estimations are made with considerable accuracy, and this is validated well enough by the results from the test dataset (the differences in the training data and test data R^2 values were not considered so large as to be a cause for concern).

Table 9-10: R^2 for manhole distributions using training and test datasets.

Land Use Category	Area Size (ha)	R^2	
		Training Data	Test Data
Residential	0-20	0.97	0.74
	20-50	0.97	0.84
	50-450	0.98	0.95
Non-Residential	0-30	0.95	0.77
	30-60	0.98	0.88
	60-160	0.98	0.89

9.3.2.2 Mean absolute percentage error (MAPE) for manhole distribution accuracy

The MAPE and 90% MAPE for the total number of manholes in each sample network are presented in Table 9-11. These results show that the MAPE for the different categories is in the order of 10% to 20%, and in the 90% best cases it is in the order of 8% to 15%. The relatively low MAPE values indicate a fairly high prediction accuracy. Interestingly, the prediction accuracy was better for large areas than for small areas. The results are validated by the results from the test dataset also presented in Table 9-11.

Table 9-11: MAPE and 90% MAPE for manhole distributions using training and test datasets.

Land Use Category	Area Size (ha)	MAPE (%)		90% MAPE (%)	
		Training Data	Test Data	Training Data	Test Data
Residential	0-20	17.3	18.7	13.5	13.1
	20-50	16.7	14.3	12.6	11.1
	50-450	15.4	14.6	9.6	11.6
Non-Residential	0-30	19.3	26.8	15.3	19.6
	30-60	11.5	10.7	8.4	6.5
	60-160	11.5	8.9	8.7	6.8

Overall, the high R^2 and low MAPE performance results suggest that the number of manholes is predicted with a reasonably high level of accuracy.

9.3.3 Manhole distribution interpretation

Only a simple estimate of the number of manholes per kilometre of pipeline was expected, but interestingly, it was found that the number of manholes per kilometre is influenced by the land use and area size. On average, there are about 20 manholes per kilometre of sewer pipeline. Predominantly residential areas tend to have slightly more manholes per kilometre, and predominantly non-residential areas tend to have slightly fewer manholes per kilometre. This is a logical outcome, since land use affects the network layout and density, which in turn affects the number of pipe junctions, and therefore, manholes. Another trend in the manhole distribution table is that, as the area size increases, the number of manholes per kilometre of pipeline tends to decrease. This is also a logical outcome, since larger areas have more large-diameter pipes, along which the maximum distance between manholes is normally increased.

9.3.4 Manhole distribution summary

The final manhole distributions were able to predict the total number of manholes in a network with very high R^2 values and a MAPE generally in the order of 20% or less. Therefore, the outcome was considered reliable.

9.4 Results Concluding Summary

In this chapter, the results for Study Outcomes I, II and III were presented. Using the three estimation tools in combination, it is possible to estimate the total sewer pipeline length per approximate diameter and the expected number of manholes associated with a service zone with a reasonable degree of confidence. The only required input characteristics of the service zone are: the dominant land use (in terms of PDDWF contribution, or alternatively, UH contribution), area size, mean elevation of the service zone, expected elevation of the network endpoint (the lowest convergence point of the network), and the number of unit hydrographs to be serviced by the network. It can therefore be concluded that the main aim of this study, which was "to develop a method for estimating the sewer pipeline infrastructure required for a service zone, based on limited information, that can be applied to both existing and future developments", was satisfactorily achieved.

An application example of the combined infrastructure estimation tool is provided in Appendix K.

Chapter 10
CONCLUSION

10.1 Overview

The aim of this project was to develop a method for estimating the gravity sewer pipeline infrastructure required for a service zone, based on limited information, that could be applied to both existing and future developments. Several studies in this vein were found in the literature. However, to date, no widely available tool was found that could be applied to reliably estimate the expected sewer pipeline infrastructure associated with a service zone in South Africa.

In order to develop a new tool for estimating sewer pipeline infrastructure in South Africa, a study was initiated comprising three major study outcomes. Study Outcome I involved the development of a model for estimating the total sewer pipeline length for a service zone using basic service zone characteristics. Study Outcome II encompassed the development of pipeline diameter distributions for disaggregating the total pipeline length into lengths per diameter, based on the characteristics of the service zone. And Study Outcome III involved quantifying the typical number of manholes required for a given length of pipeline, based on the characteristics of the service zone.

The overall approach followed during the study was to first identify the potentially influential but easily accessible service zone characteristics, through a review of sewer network design principles and the surrounding literature. The following step included sourcing a suitable South African sewer network database, from which a set of sample networks and their and relevant characteristics of interest could be extracted. Thereafter, the appropriate statistical methods were used to develop the required study outcomes. As a final step, the performance of the estimation tools arising from the study outcomes was evaluated.

A suitable database was obtained in the form of operational sewer network models for five South African municipalities, provided by GLS Consulting. From this, a dataset of 473 sample networks was extracted. The required study outcomes were successfully developed considering different land use, area size and variable availability categories.

10.2 Findings

First considering Study Outcome I, the pipeline length estimation models were compiled using weighted least squares multiple linear regression. For nine different categories of land use and area size, regression models were developed that enabled estimation of the total pipeline length as a function of three variables, namely the service zone area size, relief (in terms of mean relief, the difference between the mean elevation and the expected elevation of the network endpoint), and the density of contributing users (in terms of the number of unit hydrographs per hectare). The model performances ranged from very good to moderate, with average percentage errors in the order of 10 – 35%. Confidence intervals were also provided to allow for more conservative estimates with the poorer-performing models. A technicality of the model-estimated total pipeline length is that the network endpoint was defined as the convergence first point receiving the full combined flow from the service zone. Consequently, it is implied that the theoretical length of pipeline between the expected location of this endpoint and the nearest collector sewer would have to be estimated separately for the specific service zone layout in question.

In the total pipeline length models, the area size was by far the most significant of the three independent variables. Consequently, models for two limited variable availability cases were developed as reduced-accuracy alternatives in cases where all the independent variables were not available. Interestingly, some of the limited variable availability models performed almost as well as the full models, whereas others performed very poorly. Overall, it was concluded that the models do enable relatively accurate estimation of the total pipeline length to be made, based on varying degrees of limited information.

Regarding Study Outcome II, it was found that the pipeline diameter distributions were influenced mostly by land use, area size, and sometimes mean relief. The final product was a set of the average pipeline diameter distributions for 17 different categories of land use, area size, and mean relief. The most important outcome was that the proportion of small pipes (≤ 160 mm diameter) was always at least 90% for predominantly residential areas, and always at least 70%

for predominantly non-residential areas. The results for the remaining proportion of large-diameter pipes were somewhat more random, which suggested that the results should be applied with a conservative approach, although the average was captured fairly well.

And lastly, regarding Study Outcome III, it was found that on average, there were about 20 manholes per kilometre of sewer pipeline. However, this was slightly influenced by the land use and area size. Predominantly residential areas tended to have a slightly higher manhole frequency, and predominantly non-residential areas tended to have a slightly lower manhole frequency. Furthermore, the manhole frequency decreased slightly with increasing area size. Therefore, the number of manholes per kilometre of pipeline was tailored for six different categories of land use and area size, which allowed for reasonably accurate estimations to be made.

10.3 Limitations

This study only addressed gravity sewer infrastructure. The presence of pumps and rising mains was considered too dependent on the specific site layout to predict statistically. Nonetheless, the results may still be successfully applied to a gravity sub-catchment upstream of a rising main.

It is important to note that the results and conclusions of this study are only applicable to service zones that fall within the area size range of the samples used to develop the study outcomes. The relevant size range is specified for each of the model outcomes, but is never greater than 450 hectares. Furthermore, this tool was developed using data samples on a single-development to large-suburb scale, so the results and applicability are reflective of this.

10.4 Recommendations

There are two categories of recommendations arising from this study, namely recommendations regarding application of the study findings, and recommendations for future research.

The following recommendations are made regarding application of the study findings:

- It is recommended that future users of the infrastructure estimation tool carefully consult the relevant discussions of the data collection methods, as well as the application example in Appendix K, to ensure that the model input variables are quantified in the same way (for example, how the UHs were assigned and how the dominant land use for a service zone was determined).
- If the service zone for which the sewer pipeline infrastructure is being estimated falls near the boundary between two categories (for example, highly mixed land use with a low dominance of a single land use category), it is recommended that the relevant models and distributions for *both* of the potentially-applicable categories are implemented, and that the results are compared.
- With regard to estimation of the sewer infrastructure costs for a service zone, there exists a comprehensive cost estimation framework for sewers developed by the South African Department of Water and Sanitation or DWS (PULA, 2016), based on population size. The DWS costing tool relies upon an estimate of the pipeline length per material. Considering that most small-diameter pipes are made of PVC and most large-diameter pipes are made of concrete, the pipeline diameter distributions developed in this study could be used to improve the estimations made using the DWS costing tool.

The following recommendations are made regarding future research opportunities:

- Considering the sewer infrastructure tool developed in this study, and the equivalent tool developed for water supply infrastructure by Grotepass (2020), there are now South African methods that enable both the sewer and water supply infrastructure associated with a service zone to be estimated. This leaves room for an equivalent tool to be developed for stormwater infrastructure.
- While the sewer infrastructure tool developed in this study focussed on South African service zones of 0 – 450 hectares, the same methodology could be applied using an appropriate dataset to generate similar results that are more applicable to a specific region, land use or end application.

10.5 Closing Comments

Combined, the three study outcomes form an infrastructure estimation tool that enables reasonably accurate estimation of the sewer pipeline length per diameter and the number of manholes associated with a service zone, requiring only the following input information:

- Area size
- Land use
- DEM data for the development site.

However, it is acknowledged that there will always be project-specific variation, which cannot be accounted for statistically. As is the case in all of statistics, the true values will always deviate from the average.

Chapter 11
REFERENCES

Balaji, B., Mariappan, P. & Senthamilkumar, S., 2015. A cost estimate model for sewerage system. *ARPN Journal of Engineering and Applied Sciences,* 10(8), pp. 3327-3332.

Blumensaat, F., Wolfram, M. & Krebs, P., 2012. Sewer model development under minimum data requirements. *Environmental Earth Sciences,* Volume 65, pp. 1427-1437.

Dames & Moore, 1978. *Construction costs for municipal wastewater conveyance sytems: 1973 - 1977,* Washington, D.C.: Environmental Protection Agency.

De Veaux, R., Velleman, P. & Bock, D., 2011. Chapter 30 Multiple Regression. In: *Stats: Data and Models, 3rd Edition.* Boston: Pearson, pp. 30.1 - 30.23.

De Villiers, N., Van Rooyen, G. & Middendorf, M., 2018. Sewer network design layout optimisation using ant colony algorithms. *Journal of the South African Institution of Civil Engineering,* 60(3), pp. 2-15.

DHS, 2019. Part II Section K: Sanitation. In: *The Neighbourhood Planning and Design Guide.* Pretoria: South African Government, pp. K.1 - K.104.

Frost, J., 2020a. *Heteroscedasticity in Regression Analysis*. [Online]
Available at: https://statisticsbyjim.com/regression/heteroscedasticity-regression/#:~:text=Pure%20heteroscedasticity%20refers%20to%20cases,causes%20the%20non%2Dconstant%20variance.
[Accessed 21 June 2020].

Frost, J., 2020b. *Guidelines for Removing and Handling Outliers in Data.* [Online]
Available at: https://statisticsbyjim.com/basics/remove-outliers/
[Accessed 25 June 2020].

Ghosh, I., Hellweger, F. & Fritch, T., 2006. *Fractal Generation of Artificial Sewer Networks for Hydrologic Simulations.* San Diego, ESRI.

GLS Consulting, 2019. *Sewsan 6 Documentation,* Stellenbosch: GLS Consulting (PTY) Ltd.

GLS Software, 2020. *Sewsan - Sewer System Analysis.* [Online] Available at: https://www.gls.co.za/software/products/sewsan.html [Accessed 17 February 2020].

Greene, R., Agbenowosi, N. & Loganathan, G., 1999. GIS-Based Approach to Sewer System Design. *Journal of Surveying Engineering,* 125(1).

Grotepass, F., 2020. *A study to determine the relationship between the maximum capacity of a water reticulation network and its physical characteristics,* Stellenbosch: Stellenbosch University Department of Civil Engineering.

Haile, M., 2009. *GIS-Based Estimation of Sewer Properties from Urban Surface Information,* Dresden: Technical University of Dresden Institute of Hydrology and Meteorology.

Heaney, J., Sample, D. & Wright, L., 1999. Cost analysis and financing of urban water infrastructure. In: J. Heaney, R. Pitt & R. Field, eds. *Innovative urban wet-weather flow management systems.* Washington, DC: United States Environmental Protection Agency, pp. 10-3 - 10-25.

Jenkins, D. & Quintana-Ascencio, P., 2020. A solution to minimum sample size for regressions. *PLoS ONE,* 15(2).

Juuti, P., Maki, H. & Wall, K., 2007. Water Supply in the Cape Settlement from the Mid-17th to the Mid-19th Centuries. In: P. Juuti, T. Katko & H. Vuorien, eds. *Environmental History of Water - Global views on community water supply and sanitation.* London: IWA Publishing, pp. 165-172.

Kobayashi, T., Yamazaki, F. & Nagata, S., 2011. *Estimation of the distribution of water-pipeline length based on other infrastructure data,* Chiba: Chiba University.

Lofrano, G. & Brown, J., 2010. Wastewater Management through the Ages: A History of Mankind. *Science of the Total Environment,* 408(22), pp. 5254-5264.

Mäki, H., 2007. Developments of the Supply and Acquisition of Water in South African Towns in 1850-1920. In: P. Juuti, T. Katko & H. Vuorinen, eds. *Environmental History of Water - Global views on community water supply and sanitation.* London: IWA, pp. 173-195.

Mäki, H., 2010. Comparing developments in water supply, sanitation and environmental health in four South African cities, 1840-1920. *Historia,* 55(1), pp. 90-109.

Maurer, M., Scheidigger, A. & Herlyn, A., 2013. Quantifying costs and lengths of urban drainage systems with a simple static sewer infrastructure model. *Urban Water Journal,* 10(4), pp. 268-280.

Melton, M., 1965. The geomorphic and paleoclimatic significance of alluvial deposits in Southern Arizona. *The Journal of Geology,* 73(1), pp. 1-38.

Miller, V., 1953. *A quantitative geomorphic study of drainage basin characteristics in the Clinch Mountain area, Virginia and Tennessee; Project NR 389042, Tech. Rept. 3,* New York: Columbia University Department of Geology.

Möderl, M., Butler, D. & Rauch, W., 2009. A stochastic approach for automatic generation of urban drainage sytems. *Water Science & Technology,* 59(6), pp. 1137-1143.

Montgomery, D. C. & Runger, G. C., 2014. *Applied Statistics and Probability for Engineers.* 6th ed. Singapore: John Wiley & Sons.

Mostellar, F. & Tukey, J., 1977. *Data Analysis and Regression: A second course in statistics.* Reading: Mass Addison-Wesley Pub. Co.

NIST/SEMATECH, 2013. *e-Handbook of Statistical Methods 1.3.3.2.1. Normal Probability Plot.* [Online]
Available at:
https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm#:~:text=The%20normal%20probability%20plot%20(Chambers,form%20an%20approximate%20straight%20line.
[Accessed 29 September 2020].

Pauliuk, S., Venkatesh, G., Brattebo, H. & Muller, D., 2014. Exploring urban mines: pipe length and material stocks in urban water and wastewater networks. *Urban Water Journal,* 11(4), pp. 274-283.

Pennsylvania State University, 2018. *STAT 462 10.1 - Nonconstant Variance and Weighted Least Squares.* [Online]

Available at: https://online.stat.psu.edu/stat462/node/186/
[Accessed 10 May 2020].

PULA, 2016. *Cost benchmark for water services projects,* Pretoria: Department of Water and Sanitation.

Roscoe, J., 1975. *Fundamental Research Statistics for the Behavioral Sciences.* New York: Holt, Rinehart and Winston.

SAICE, 2017. *SAICE 2017 Infrastructure Report Card for South Africa,* Midrand: South African Institution of Civil Engineering.

Schumm, S., 1956. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *Geological Society of America Bulletin,* 67(5), pp. 597-646.

Shlens, J., 2014. A tutorial on principal component analysis. *International Journal of Remote Sensing,* Volume 2, pp. 1-13.

Sitzenfrei, R., Fach, S., Kinzel, H. & Rauch, W., 2010a. A multi-layer cellular automata approach for algorithmic generation of virtual case studies - VIBe. *Water Science & Technology,* 61(1), pp. 37-45.

Sitzenfrei, R. et al., 2010b. Dynamic virtual infrastructure benchmarking: DynaVIBe. *Water Science & Technology,* 10(4), pp. 600-608.

South Africa, 1997. *Water Services Act 108 of 1997.* [Online]
Available at: https://www.gov.za/sites/default/files/gcis_document/201409/a108-97.pdf
[Accessed 12 February 2020].

Statistics South Africa, 2018. *General Household Survey,* Pretoria: Statistics South Africa.

Stephenson, D. & Barta, B., 2005. *Guidelines on reduction of the impact of water infiltration into sewers,* Gezina: Water Research Commission.

Strahler, A., 1964. Quantitative geomorphology of drainage basins and channel networks. In: V. Chow, ed. *Handbook of Applied Hydrology.* New York: McGraw Hill, pp. 439-476.

Swamee, P., 2001. Design of Sewer Line. *Journal of Environmental Engineering,* 127((9)), pp. 776-781.

Urich, C., Sitzenfrei, R., Möderl, M. & Rauch, W., 2010. An agent-based approach for generating virtual sewer systems. *Water Science & Technology,* 62(5), pp. 1090-1097.

van Vuuren, S. & van Dijk, M., 2011. *Waterborne Sanitation Design Guide,* Gezina: Water Research Commission.

Wilson, J. & Gallant, J., 2000. *Terrain Analysis: Priciples and Applications.* New York: John Wiley & Sons.

WWF-SA, 2016. *Water Facts and Futures: Rethinking South Africa's Water Future,* Cape Town: WWF-SA.

Yen, B. & Sevuk, A., 1975. Design of Storm Sewer Networks. *Journal of Environmental Engineering Division,* 101(EE4), pp. 535-553.

Yen, B., Wenzel, H., Mays, L. & Tang, W., 1976. *Advanced Methodologies for Design of Storm Sewer System,* Urbana: Water Resources Center, University of Illinois.

Zavoianu, I., 1978. *Morphometry of Hydrographic Basins.* Bucharest: Editura Academiei.

Zavoianu, I., 1985. *Morphometry of Drainage Basins (Developments in Water Science, 20).* 2 ed. Bucharest: Editura Academiei.

PART 2 - APPENDICES

Appendix A

SUPPORTING INFORMATION FOR DWS SEWER COST BENCHMARK

Table A-1 provides the adjustment factors that can be applied to the cost estimates from the DWS sewer project costing guide (PULA, 2016), in order to account for project-specific variations.

Table A-1: Adjustment factors DWS sewer project cost benchmark (PULA, 2016).

Factors Affecting Costs		Bulk pipeline	Reticulation
Project size	<1500 people	+5%	+5%
	1500 – 5000 people	0	0
	>5000 people	-3%	-3%
Distance from economic centre	<50 km	0	0
	50 – 100 km	+10%	+10%
	>200 km	+15%	+15%
Topography	<1% slope	0	+2%
	1 – 5% slope	0	0
	>5% slope	+5%	+5%
Clearing	Savannah	0	-
	Bush	+2%	-
	Trees	+5%	-
Availability of contractor	High	-10%	-5%
	Medium	0	0
	Low	+15%	+10%
Geology	Soft	0	0
	Intermediate	+30%	+30%
	Hard rock	+60%	+60%
Land acquisition & servitudes	Public area	0	-
	Agricultural land	+1%	-
	Built-up area	+4%	-

Appendix B
METHODS FOR QUANTIFYING THE FORM CHARACTERISTICS OF A SERVICE ZONE

Among the service zone characteristics of interest that were identified as potentially-influential factors in the development of the three study outcomes, the inherent physical characteristics of the land itself were encompassed. These so-called 'form characteristics' comprised the area size, shape, slope and topography. In this appendix, different methods or indicators for quantifying the form characteristics are investigated and discussed.

Many of the form indicators discussed in this appendix were obtained from the literature, and largely from the field of hydrological geomorphometry (discussed in Section 2.5.3). It is noted that most of the evidence supporting the efficacy of the different indicators is empirical rather than theoretical, and it is generally difficult to distinguish a 'best' indicator representing each form characteristic. Furthermore, since hydrological geomorphometry concerns large-scale river catchments, it was considered possible that many of the geomorphometric indicators might not translate well to the smaller-scale sewer 'catchments' considered in this study. Therefore, for each form characteristic, multiple quantification methods or indicators were investigated, from which the most promising few were selected for consideration as candidate variables in the development of the study outcomes. In the following five sections, quantification indicators are discussed for the five major form characteristics of interest for this study, namely area size, area length, area shape, slope, and topography. All of the indicators discussed are summarised in a table at the end, followed by a list of references.

B.1 Area Size

The *real* surface area of a study area can be calculated relatively accurately by summing the areas of triangles formed between the points of a digital elevation model (DEM) of a study area. Alternatively, the true area size can be approximated by the *plane* area, or the horizontal projection of the real surface area, which is more easily calculated. However, the plane area can deviate substantially from the real surface area in study areas where the slope or hilliness is significant (Zavoianu, 1985).

B.2 Area Length

The length of a study area provides some indication of the size and shape of a study area. However, since this property is not clearly defined and its quantification methods are approximate, Zavoianu (1985) cautioned that use of this property may reduce the accuracy of a model. Nonetheless, Zavoianu (1985) described two main approaches to quantifying the area length in the context of river catchments, namely mouth-source methods and the average length.

B.2.1 Mouth-source line

In hydrological geomorphometry, many researchers have defined catchment length along some line drawn through the catchment according to certain criteria (Zavoianu, 1985). Horton (1932) described a straight line from the catchment mouth to the point where the catchment perimeter is intersected by the projection of the main stream source. Apollov (1963) described a straight line from the catchment mouth to the furthest-away point on the perimeter; Ogievsky (1952) described a similar line with the same endpoints but which passes through the midpoints of lines drawn across the catchment. Overall, manual methods of this type were not considered to be practical for implementation in this study and were given no further consideration.

B.2.2 Average length

If a catchment is approximated as a rectangle, with length and breadth dimensions $\bar{L}$ and $\bar{B}$ such that $A = \bar{L} \times \bar{B}$ and $P = 2(\bar{L} + \bar{B})$, then expressions can be derived to represent $\bar{L}$ and $\bar{B}$ in terms of the measured area (A) and perimeter (P). Equations B-1 and B-2 (Zavoianu, 1985) thus give the average length and average breadth of the catchment as a function of A and P. For cases when $\mathrm{A} > (\mathrm{P} \div 4)^2$, which can occur in circular catchments, Equation B-3 (Zavoianu, 1985) may be used to calculate average length in terms of A and P. Although an approximation, Zavoianu (1985) found that this measure of average length generally showed a good correlation with the main stream length in a catchment, and generally gave good results when used with certain shape factors (discussed in Appendix B.3).

$$\bar{L} = \frac{P}{4} + \sqrt{(\frac{P}{4})^2 - A} \quad \text{B-1}$$

$$\bar{B} = \frac{P}{4} - \sqrt{(\frac{P}{4})^2 - A} \quad \text{B-2}$$

$$\bar{L} = 4(\frac{A}{P}) \quad \text{B-3}$$

A general limitation of any shape indicators that rely on the area and perimeter is that they may become less accurate with irregularities in the shape boundary. For example, the perimeter is affected more by indents or outcrops than the area is, which could skew the perimeter-area relationship and therefore misrepresent the shape of the study area.

B.3 Area Shape

In spite of the potential inaccuracies of perimeter-area indicators, the relationship between the size of a study area and its perimeter can be valuable in describing how elongated the study area shape is. In the following sub-sections, six shape factors are discussed, namely the circularity ratio, compactness coefficient, Horton's form factor, the elongation ratio, Zavoianu's form factor, and centroid-mouth relative radius.

B.3.1 Circularity ratio

The circularity ratio proposed by Miller (1953) is defined as the ratio of the study area size (A), to the size of a circle with a circumference equal to the perimeter (P) of the study area. Simplified, the circularity ratio is defined in Equation B-4. The ratio has a maximum value of one for a circular study area, a value of 0.785 for a square, and continues to decrease as the study area becomes more elongated.

$$Circularity\ Ratio = \frac{4\pi A}{P^2} \quad \text{B-4}$$

B.3.2 Compactness coefficient

The compactness coefficient relates the perimeter (P) of the study area to the perimeter of a circle with the same area size (A) (Luchisheva, 1950). Simplified, the compactness coefficient is defined in Equation B-5.The ratio has a minimum value of one when the study area is circular, a value of 1.128 when the shape is square, and increases as the shape becomes more elongated.

$$Compactness\ Coefficient = \frac{0.282\ P}{\sqrt{A}} \qquad \text{B-5}$$

B.3.3 Horton's form factor

The form factor proposed by Horton (1932) represents the study area size (A) divided by the maximum length of the study area. Zavoianu (1985) found that if the maximum length is replaced with the average length $\bar{L}$ described in Equation B-1, then calculated values for Horton's form factor correlate strongly with calculated values of the circularity ratio. Thus simplified, Horton's form factor is defined in Equation B-6. The form factor has a value of 1.273 if the study area is circular, a value of one if the study area is square, and decreases with increasing elongation.

$$Horton's\ Form\ Factor = \frac{A}{\bar{L}^2} \qquad \text{B-6}$$

B.3.4 Elongation ratio

In the context of catchments, the elongation ratio is defined as the diameter of a circle of equal area to the study area (A), divided by the maximum catchment length measured parallel to the main stream axis (Schumm, 1956). Similarly to Horton's form factor, if the maximum catchment length is replaced with the average length $\bar{L}$ (Equation B-1), then the elongation ratio correlates better with the circularity ratio (Zavoianu, 1985). Thus simplified, the elongation ratio is defined in Equation B-7 (Seyhan, 1977). The value of the elongation ratio is one for a circular study area and decreases with increasing elongation. According to Zavoianu (1985), research suggests that rainfall and runoff are correlated better with the elongation ratio than Horton's form factor, which may suggest that it is a better shape indicator.

$$Elongation\ Ratio = \frac{1.129\sqrt{A}}{\bar{L}} \quad \text{B-7}$$

B.3.5 Zavoianu's form factor

Since catchments are never circular in reality, Chorley et al. (1957) suggested that comparing catchment shape to a square may be more appropriate. Zavoianu (1978) proposed a form factor that uses as a reference shape a square with perimeter (P) equal to that of the study area. The area of the reference square is given by $(\frac{P}{4})^2$. Zavoianu's form factor is then the ratio of the study area size (A), to that of the reference square. Simplified, Zavoianu's form factor is defined in Equation B-8. Zavoianu's form factor has a value of one for square study areas, a value greater than one for study areas approaching a circular shape, and a value less than one for a more elongated shape. Zavoianu's form factor shows a good correlation with other shape factors (Zavoianu, 1985), and is directly related to the circularity ratio as shown in Equation B-9.

$$Zavoianu's\ Form\ Factor = \frac{16A}{P^2} \quad \text{B-8}$$

$$Circularity\ Ratio = 0784\ \times Zavoianu's\ Form\ Factor \quad \text{B-9}$$

B.3.6 Centroid-mouth relative radius

To give an indication of the shape of the sewer network, the centroid-mouth relative radius was defined for this study simply as the plane distance between the study area centroid (coordinates X_c:Y_c) and the end manhole or network mouth (coordinates X_{mouth}:Y_{mouth}), normalised by the square root of the study area size (A). The centroid-mouth relative radius is defined in Equation B-10.

$$Centroid\text{-}Mouth\ Relative\ Radius = \frac{\sqrt{(X_c - X_{mouth})^2 + (Y_c - Y_{mouth})^2}}{\sqrt{A}} \quad \text{B-10}$$

B.4 Slope

In the context of river catchments, slope is normally not constant but varies over the catchment. Slope is therefore typically defined as a mean slope, or the mean slope of a specific line through the catchment. According to Stephenson and Barta (2005), the slope near the centre of a catchment is generally less than 6%, and often less than 3%.

In hydrological geomorphometry, a fairly accurate measurement is to find the slope between the contours and calculate the mean (Zavoianu, 1985). Another popular measure is the mean slope of the main stream. But, in the context of sewer 'catchments', owing to the limited data that would normally be available for slope calculations for a greenfield project, the slope indicators described in this section are more approximate in nature. In the following sub-sections, two slope factors are discussed, namely mean basin slope and mean perimeter slope.

B.4.1 Mean basin slope

In the context of catchments, the mean basin slope is the difference between the highest (H_{max}) and lowest (H_{min}) points in a catchment, divided by the length of the catchment (Schumm, 1956). The length of the catchment can be approximated by $\bar{L}$ in Equation B-1, yielding Equation B-11. While this does not necessarily signify the true catchment slope, it may still be useful approximation. According to Zavoianu (1985), in general, the maximum catchment altitude and the altitude of the source of the main river do not differ greatly. Therefore, for catchments of relatively homogeneous relief, there tends to be a good correlation between the mean basin slope and the slope of the main stream. However, this relationship is less applicable in regions of less homogeneous relief.

$$Mean\ Basin\ Slope = \frac{H_{max} - H_{min}}{\bar{L}} \qquad \text{B-11}$$

B.4.2 Mean perimeter slope

In the context of catchments, the mean perimeter slope is the maximum elevation on the water divide (H_{max}) minus the elevation of the river mouth (H_{mouth}), divided by half of the perimeter length (P), as defined in Equation B-12 (Zavoianu, 1985). The mean perimeter slope normally

gives values less than but quite close to the mean slope of the main channel. It is more accurate when the perimeter has a homogeneous slope and becomes less accurate when there are saddles and peaks, as are more likely to occur in larger catchments. For highly asymmetrical catchments, it is more appropriate to calculate the mean perimeter slope separately for each side.

$$Mean\ Perimeter\ Slope = \frac{2(H_{max} - H_{mouth})}{P} \quad \text{B-12}$$

B.5 Topography

There are a number of indicators that have been used to describe the relief and hilliness of terrain, sometimes referred to as topographical ruggedness or roughness. According to Pierce and Kolden (2015), such indicators were developed for varying applications, from measuring incident solar radiation to describing the habitats of animals. But, to the aforementioned authors' knowledge, there has not been any systematic attempt to identify the best indicators for different applications. Therefore several indicators are discussed in the following sub-sections, namely total relief, mean relief, deviation from mean elevation, elevation standard deviation, the ruggedness number, Melton's ruggedness number, and the surface area ratio. Most of these indicators require as inputs either a dataset of the elevation at different points in the study area; or the heights of the river mouth, the highest point, and the lowest point; or all of these.

B.5.1 Total relief

The total relief or elevation range is the difference in elevation between the highest (H_{max}) and lowest (H_{min}) points in the area, as defined by Equation B-13 (Zavoianu, 1985). A disadvantage of this indicator is that it may be skewed or change abruptly depending on whether local high or low points are included in the study area (Wilson & Gallant, 2000).

$$Total\ Relief = H_{max} - H_{min} \quad \text{B-13}$$

B.5.2 Mean relief

In the context of catchments, mean relief describes the elevation of the river endpoint (H_{mouth}) relative to the mean elevation of the catchment (H_{mean}), given in Equation B-14 (Wilson & Gallant, 2000). The mean relief is often used for modelling processes that are affected by local differences from the overall elevation, such as groundwater flow (Wilson & Gallant, 2000).

$$Mean\ Relief = H_{mean} - H_{mouth} \quad \text{B-14}$$

B.5.3 Elevation standard deviation

The standard deviation of the elevation is determined by applying Equation B-15 (Montgomery & Runger, 2014) to a dataset of the elevation at different points in the study area, such as in a DEM. The elevation standard deviation indicates the variability of elevation within the study area. According to Wilson and Gallant (2000), for small study areas, the elevation standard deviation indicates the local relief, and as study areas become larger and incorporate more local high and low points, it indicates the roughness of the landscape.

$$Elevation\ Standard\ Deviation = \sqrt{\frac{\sum(H_i - H_{mean})^2}{N}} \quad \text{B-15}$$

Where: H_i represents each elevation value from the dataset, H_{mean} represents the mean elevation, and N represents the size of the dataset.

B.5.4 Deviation from mean elevation

In the context of catchments, deviation from mean elevation is the mean relief (Equation B-14) divided by the elevation standard deviation (Equation B-15), as defined by Equation B-16 (Wilson & Gallant, 2000). The deviation from mean elevation indicates the height of the river mouth relative to the mean elevation of the catchment, normalised to the local surface roughness. The value normally lies between -1 and 1.

$$Deviation\ from\ Mean\ Elevation = \frac{H_{mean} - H_{mouth}}{Elevation\ Standard\ Deviation} \quad \text{B-16}$$

B.5.5 Ruggedness number

In the context of catchments, the ruggedness number is defined as the product of the stream length per unit area and the total relief (Equation B-13), divided by 1000 (Strahler, 1964). The ruggedness number is defined in Equation B-17. The ruggedness number gives an indication of the structural complexity of the terrain, or the extent of instability of the land surface (Khakhlari & Nandy, 2016).

$$Ruggedness\ Number = \frac{Total\ Stream\ Length \times (H_{max} - H_{min})}{Area\ Size\ \times 1000} \qquad \text{B-17}$$

B.5.6 Melton's ruggedness number

Melton's ruggedness number is defined as the ratio of the total relief (Equation B-13) to the square root of the area size (A), as defined in Equation B-18 (Melton, 1965). It provides an indication of the relief scaled to the size of the area.

$$Melton's\ Ruggedness\ Number = \frac{H_{max} - H_{min}}{\sqrt{A}} \qquad \text{B-18}$$

B.5.7 Surface area ratio

For this study, it was considered that a simple ratio of the real surface area to the plane surface area could also provide a useful indication of the general deviation of the study area from being perfectly flat. This so-called surface area ratio is expressed in Equation B-19. The surface area ratio does not distinguish between general slope and hilliness, but rather attempts to account for the combined effect of both.

$$Surface\ Area\ Ratio = \frac{Real\ Surface\ Area}{Plane\ Surface\ Area} \qquad \text{B-19}$$

B.6 Concluding Summary

All the indicators investigated for quantifying form characteristics are summarised in Table B-1. From these indicators, some (marked with an asterisk '*') were chosen to use as candidate variables in the development of the study outcomes. When applying those indicators that were developed specifically in the context of river catchments to the context of this study, a service zone was considered as a catchment, and the associated sewer network was considered as the stream or river. Since it was unclear which indicator best described each form characteristic, multiple indicators were selected for some of the form characteristics, assuming that the best indicator would give the best results in the analysis. It is noted that no indicators to describe the area length were selected, since it was considered that the combined effect of the area size and shape adequately covered this aspect. It is also noted that the two indicators for slope (mean basin slope and mean perimeter slope) were grouped with the topography indicators.

Table B-1: Summary of form characteristic indicators investigated for use in analysis.

Form Characteristic	Indicator	Definition	Reference
Area Size	Plane area *	-	-
	Real surface area	-	-
Area Length	Mouth-source line	-	(Zavoianu, 1985)
	Average length	$\bar{L} = \frac{P}{4} + \sqrt{(\frac{P}{4})^2 - A}$ $\bar{L} = 4\left(\frac{A}{P}\right) if\ A > \left(\frac{P}{4}\right)^2$	(Zavoianu, 1985)
Area Shape	Circularity ratio *	$\frac{4\pi A}{P^2}$	(Miller, 1953)
	Compactness coefficient	$\frac{0.282\ P}{\sqrt{A}}$	(Luchisheva, 1950)
	Horton's form factor	$\frac{A}{\bar{L}^2}$	(Horton, 1932)
	Elongation ratio *	$\frac{1.129\sqrt{A}}{\bar{L}}$	(Schumm, 1956)
	Zavoianu's form factor	$\frac{16A}{P^2}$	(Zavoianu, 1978)
	Centroid-mouth relative radius *	$\frac{\sqrt{(X_c - X_{mouth})^2 + (Y_c - Y_{mouth})^2}}{\sqrt{A}}$	-
Topography	Mean basin slope *	$\frac{H_{max} - H_{min}}{\bar{L}}$	(Schumm, 1956)
	Mean perimeter slope *	$\frac{2(H_{max} - H_{mouth})}{P}$	(Zavoianu, 1985)
	Total relief *	$H_{max} - H_{min}$	(Zavoianu, 1985)
	Mean relief *	$H_{mean} - H_{mouth}$	(Wilson & Gallant, 2000)
	Elevation standard deviation *	$\sqrt{\frac{\sum(H_i - H_{mean})^2}{N}}$	-
	Deviation from mean elevation *	$\frac{H_{mean} - H_{mouth}}{Elevation\ Standard\ Deviation}$	(Wilson & Gallant, 2000)
	Ruggedness number *	$\frac{Total\ Stream\ Length \times (H_{max} - H_{min})}{A \times 1000}$	(Strahler, 1964)
	Melton's ruggedness number *	$\frac{H_{max} - H_{min}}{\sqrt{A}}$	(Melton, 1965)
	Surface area ratio *	$\frac{Real\ Surface\ Area}{Plane\ Surface\ Area}$	-

* Selected to be used as candidate variable in development of the study outcomes.

B.7 Appendix B References

Apollov, B., 1963. *A Study of Rivers*, Moscow: Izdat. Moskva Universitet.

Chorley, R., Malm, D. & Pogorzelski, H., 1957. A New Standard for Estimating Drainage Basin Shape. *American Journal of Science*, Volume 255, pp. 138-141.

Horton, R., 1932. Drainage Basin Characteristics. *Transactions of the American Geophysical Union*, 13(1), pp. 350-361.

Khakhlari, M. & Nandy, A., 2016. Morphometric Analysis of Barapani River Basin In Karbi Anglong District, Assam. *International Journal of Scientific and Research Publications*, 6(10), pp. 238-249.

Luchisheva, A., 1950. *Practical Hydrology*. Leningrad: Gidrometeoizdat.

Melton, M., 1965. The geomorphic and paleoclimatic significance of alluvial deposits in Southern Arizona. *The Journal of Geology*, 73(1), pp. 1-38.

Miller, V., 1953. *A quantitative geomorphic study of drainage basin characteristics in the Clinch Mountain area, Virginia and Tennessee; Project NR 389042, Tech. Rept. 3*, New York: Columbia University Department of Geology.

Montgomery, D. C. & Runger, G. C., 2014. *Applied Statistics and Probability for Engineers*. 6th ed. Singapore: John Wiley & Sons.

Ogievsky, A., 1952. *Land Hydrology: General and Engineering*, Moscow: Gosudarstvo Indate Seiskohoz. Literatury.

Pierce, J. & Kolden, C., 2015. The Hilliness of U.S. Cities. *Geographical Review*, 105(4), pp. 581-600.

Schumm, S., 1956. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *Geological Society of America Bulletin*, 67(5), pp. 597-646.

Seyhan, E., 1977. *The watershed as a hydrological unit: Issue 63 of Publicatie uit het Geografisch Instituut der Rijksuniversiteit te Utrecht*. Utrecht: Geografisch Instituut Utrecht.

Stephenson, D. & Barta, B., 2005. *Guidelines on reduction of the impact of water infiltration into sewers*, Gezina: Water Research Commission.

Strahler, A., 1964. Quantitative geomorphology of drainage basins and channel networks. In: V. Chow, ed. *Handbook of Applied Hydrology*. New York: McGraw Hill, pp. 439-476.

Wilson, J. & Gallant, J., 2000. *Terrain Analysis: Priciples and Applications*. New York: John Wiley & Sons.

Zavoianu, I., 1978. *Morphometry of Hydrographic Basins*. Bucharest: Editura Academiei.

Zavoianu, I., 1985. *Morphometry of Drainage Basins* (Developments in Water Science, 20). 2 ed. Bucharest: Editura Academiei.

Appendix C
MODIFICATIONS TO SAMPLE NETWORKS

Pipe diameters were modified in the sample networks for two reasons. The first reason was to ensure that all pipes in the network were operating with sufficient spare capacity. The second reason was to ensure that remnants of cut-off lines, which originally conveyed additional flow generated outside of the service zone, were resized appropriately for conveying only the flow generated within the service zone. In this addendum, the procedure followed to implement these changes is described, as well as the technical difficulties encountered, solutions to these issues, and the implications thereof.

C.1 Isolating Networks

After a suitable sample network was identified, the first step was to isolate it. Figure C-1 shows a desired sample network with a through-line conveying additional flow from an upstream service zone. It is noted that such a sample network was only considered acceptable if the presence of the through-line did not obviously alter the layout of the network from the layout that could be reasonably expected had the through-line not been there. To isolate the desired sample network, the connection immediately downstream of the sample network endpoint was removed, as was the connection upstream of the point where the additional flow entered the sample network.

After a desired sample network was isolated, all pipes downstream of the upstream cut-off point had to be resized for the decreased flow volume. This was done first by manually downsizing the affected pipes to the minimum realistic diameter (based on the surrounding pipe diameters), and then running a Sewsan planning analysis to upsize the affected pipes where necessary to ensure that all pipes in the sample network were operating with sufficient spare capacity. The manual downsizing and automated upsizing of affected pipes are described in the next two steps.

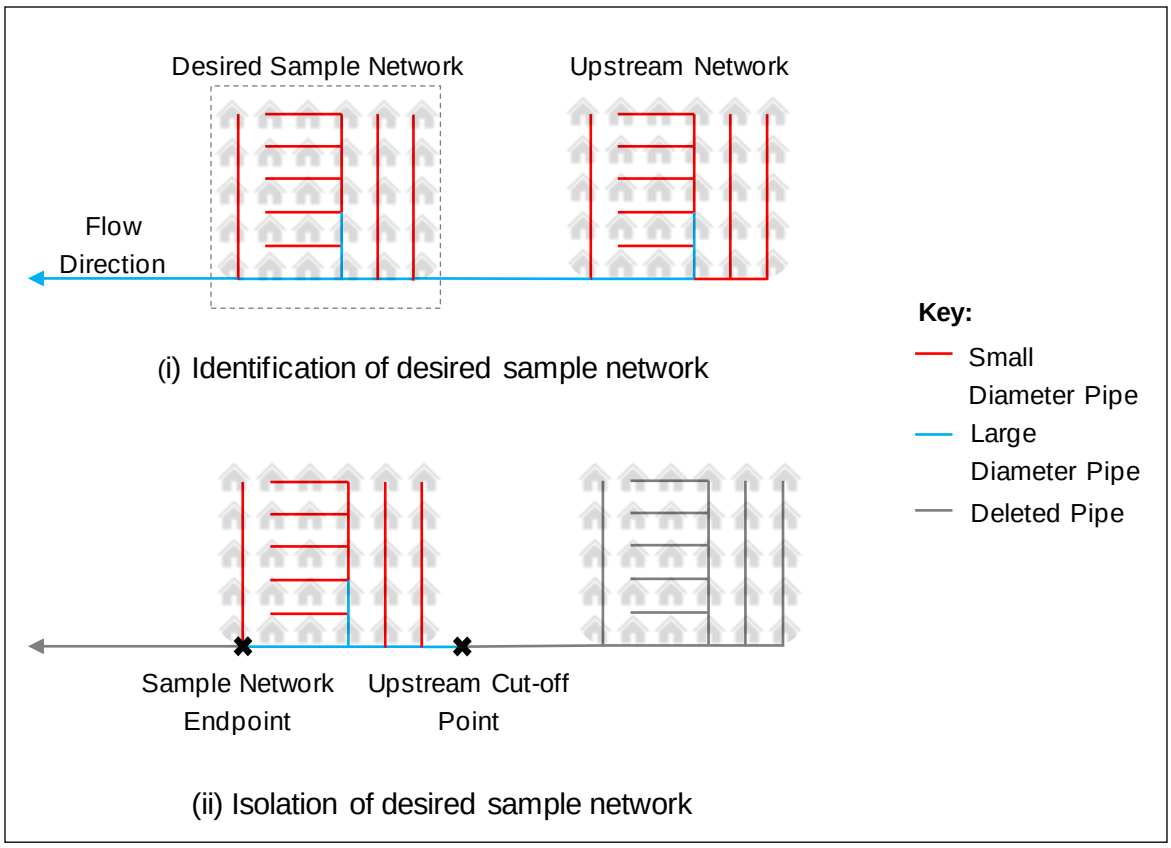


Figure C-1: Isolating a sample network from a larger network.

C.2 Downsizing Pipes

The pipe diameters were manually downsized according to a set of rules. Firstly, for pipes with no pipes upstream ('source' pipes), the diameter was set equal to that of the surrounding reticulation source pipes. In almost all cases, the diameters of the surrounding source pipes were consistent; but in cases where more than one option was available, then the smallest was selected. This is illustrated in Figure C-2 (ii).

Thereafter, moving downstream towards the sample network endpoint, at each manhole the downstream pipe was set to the same diameter as the largest pipe directly upstream of the manhole. This ensured that no pipes were downsized to be smaller than any of their upstream pipes. This is illustrated in Figure C-2 (iii).

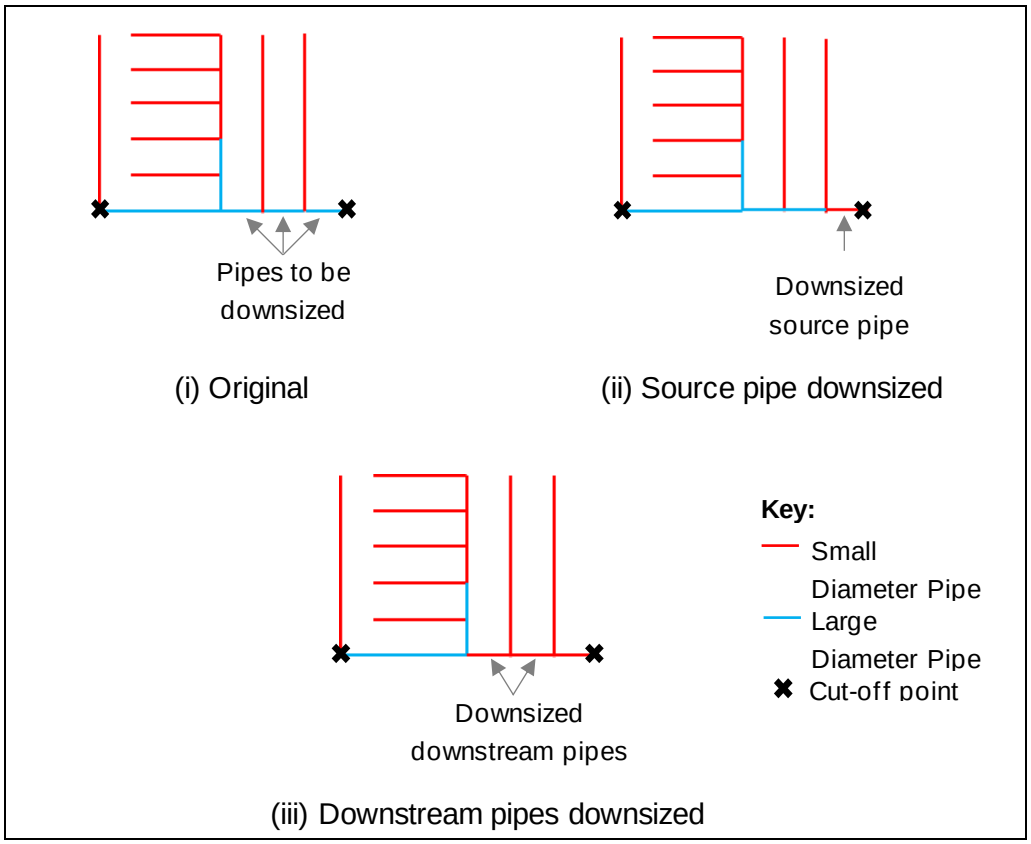


Figure C-2: Manual downsizing of pipes.

C.3 Upsizing Pipes

To ensure that all pipes in the system were operating with a sufficient spare capacity of at least 30%, a Sewsan planning analysis was run. This addressed both the unaltered pipes designed with insufficient capacity, as well as the modified pipes on the through-line. This is illustrated in Figure C-3, where one of the manually-downsized pipes was too narrow, and had its diameter increased again.

As mentioned in the literature review, the planning analysis function iteratively increases the diameters of pipes of insufficient capacity until there are no more bottlenecks and the user-defined spare capacity is accommodated. The new diameters in the planning analysis were limited to those contained in Table C-1.

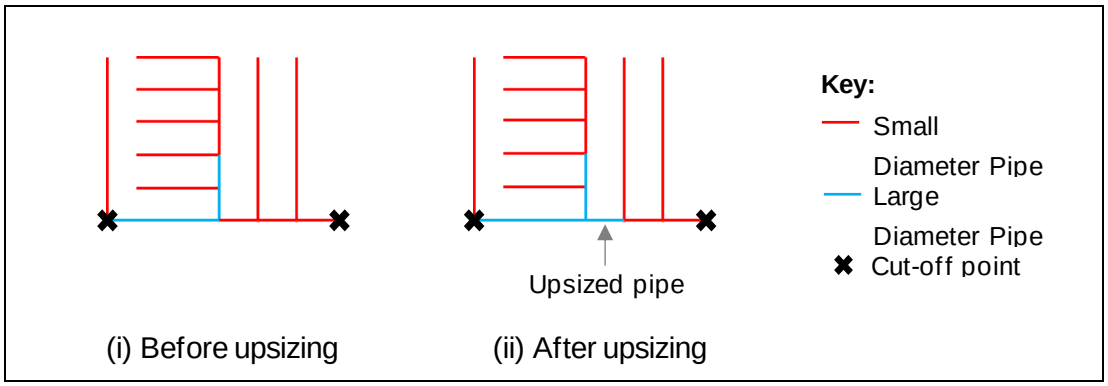


Figure C-3: Automated upsizing of pipes.

Table C-1: Allowed new internal and nominal pipe diameters.

Diameter (mm)			
Internal	**Nominal**	**Internal**	**Nominal**
104	110	633	675
151	160	704	750
188	200	762	825
235	250	843	900
297	315	1008	1050
335	355	1149	1200
377	400	1290	1350
419	450	1423	1500
488	525	1602	1650
559	600	1717	1800

C.4 Random Large Diameters and Slope Adjustments

A technical issue encountered when running the Sewsan planning analysis to upsize pipes with insufficient capacity, was the occasional and seemingly-random occurrence of very large diameter pipes on the modified through-lines. This was caused by the pipe slopes.

Larger pipes are generally laid at flatter slopes to ensure maximum velocity requirements are not exceeded – consequently, pipes with large original pipe diameters in the models also tended to have flatter slopes. The inverse is also true, that pipes with fixed flat slopes generally require larger diameters to ensure minimum velocity requirements are met. As such, since the new

diameter was a function of the slope, and the slope was a function of the original diameter, then the new diameter could be influenced by the original diameter. This only had an effect in the pipes where the original slope was flatter than the minimum slope of the new diameter. In such cases, the pipe was immediately upsized by Sewsan to a diameter for which the actual slope was not flatter than the minimum slope for this diameter, thus producing the random large-diameter pipes.

In Sewsan models, pipe slope can be specified in the different ways specified in Table C-2. The very flat slopes arose in two situations. Firstly, if the slope type was 'Minimum', this would automatically be changed to 'Fixed Minimum' at start of planning analysis, thus constraining it to the minimum slope for the original diameter. This was done on the assumption that the slope of an existing pipe was fixed. However, the 'Minimum' slope type did not represent the actual pipe slope, but only an estimate. Therefore, it was considered that changing the slope to the minimum slope for the new diameter instead would be an acceptable change, as no real slope information would be lost. This was achieved by modelling the pipes as not-yet existing, thus preventing the automatic conversion to a 'Fixed Minimum' slope type. This accounted for most of the random large-diameter pipes,

Table C-2: Sewsan gravity pipe slope types (GLS Consulting, 2019).

Slope Type	Description
Inverts	Slope is calculated using invert levels at the pipe ends, and the pipe length, provided the upstream invert is lower than the downstream invert of the previous pipe.
Datum	Slope is calculated using the invert levels at the pipe ends, and the pipe length.
Slope	User-defined slope.
Minimum	Minimum slope allowed for the pipe diameter.
Fixed Minimum	Minimum slope allowed for the original pipe diameter.
Estimate	Estimated from invert levels of its upstream and downstream pipes.

The second cause for very flat pipe slopes was when the actual recorded or calculated slope was very flat – this could occur with slope types 'Inverts', 'Datum', 'Slope' and 'Estimate'. The difference for these cases was that changing the slope would represent a loss of real slope information. But, since these represented very few pipes overall, it was considered that changing

the slopes of problematic pipes only would be acceptable. Therefore, the slopes were adjusted only for those pipes that had been manually downsized, whose slopes were flatter than the minimum slope for their new downsized diameters. The new slopes were set to type 'Minimum' and the pipes modelled as not-yet existing, thus constraining the new slope to the minimum slope of the resized diameter.

The implication of changing some pipe slopes in the sample network models was that the lengths of these pipes would be affected. However, for 'Minimum' slope types, the length was already based on an estimate. Therefore, the only pipes affected were the very few whose real slopes were changed. However, such small-scale slope changes as in this case had a negligible effect on the total length of a pipeline, and would not affect the accuracy of the pipeline length estimation model.

Appendix D
DATA EXTRACTION FROM SEWSAN MODELS

D.1 Defining the Bounding Polygon for a Service Zone

Each sample network required a bounding polygon defining the service zone, from which the area and shape characteristics could be derived. These polygons were drawn manually, and specific rules were followed to ensure consistency. The rules are described here to ensure repeatability.

Firstly, the erven belonging to the service zone were identified by inspection, considering the pipes and the background aerial photograph. Physical boundaries, such as fences, roads, and open space clarified the service zone borders. Figure D-1 is an example of a straightforward case. When it was unclear if an erf belonged to a service zone, the number of unit hydrographs (UHs) assigned at the network nodes were considered. Figure D-2 shows an example where some of the buildings did not appear to be connected to the network, but the UHs indicated that they were. In some cases, such as large neighbourhoods, the service zone of interest was a sub-catchment of a larger development serviced by multiple bordering sewer networks, as illustrated in Figure D-3. In these cases, more judgement was required. The distances of the erven from the different pipes were considered, and sometimes the strip of erven between two different service zones were simply separated in half. For bigger residential areas, public open spaces or undeveloped land were also often incorporated in the service zone. If these lay along the edge of the service zone, and if they clearly belonged to the area, they were fully included. If they were shared with a bordering sub-catchment, then they were divided appropriately. However, it was important that the polygon borders remained as smooth and regular as possible to prevent the area-perimeter relationship from being skewed, which would affect the shape factors.

Of course, there was inherent subjectivity in this method of service zone delineation, and the boundaries drawn might differ from user to user. However, the size of the potential errors introduced by this were considered acceptably small relative to the total area size. From each bounding polygon, the area size, perimeter length, and the coordinates of the centroid of the bounding rectangle were extracted. The centroid of bounding rectangle was used instead of the true centroid, since obtaining the centroid of an irregular shape was not supported by the software.

Figure D-1: Example of a simple service zone polygon.

Figure D-2: Example of a less obvious service zone polygon.

Figure D-3: Example of a service zone polygon directly bordered by other service zones.

D.2 Assignment of Unit Hydrographs

Table D-1 shows the guideline for the assignment of unit hydrographs (UH) for different land uses that was used to populate the source data models. The unit hydrographs were assigned for a future development scenario where empty erven that had already been zoned for development were also populated. It is noted that connections of the same land use of the same size would have the same number of unit hydrographs assigned, but not necessarily the same water demand or wastewater flow production.

Table D-1: Assignment of unit hydrographs for different land uses.

Land Use	Unit Representing One UH
Rural	erf
Low density residential	erf
Medium density residential	erf
High density residential	erf
Cluster	unit
Flats	unit
RDP[5]	unit
Informal	unit
Low cost housing	erf
Business/ commercial	100m² floor
Industrial	100m² floor
Warehousing	100m² floor
Mixed	100m² floor
Parks	ha
Densification (residential)	unit
Densification (BCI[6])	100m² floor
Educational	unit
Institutional	100m² floor
Mine	ha
Large	100m² floor
Farm/ agricultural holding	ha

[5] Reconstruction and Development Programme
[6] Business/ commercial/ institutional

D.3 Flow Definition and Calculation

D.3.1 Flow definition

Several options were available for defining the sample network flow. This included the average daily dry weather flow (ADDWF); peak daily dry weather flow (PDDWF); instantaneous peak dry weather flow (IPDWF); instantaneous peak wet weather flow (IPWWF); and the option to specify any of the above with or without infiltration and ingress. A detailed discussion on these concepts is provided in the Literature Review (Section 2.2.2). For the selected flow definition to provide a suitable solution, it had to satisfy two requirements. Firstly, the data required to calculate it would have to be available at the early phases of a project when the infrastructure estimation tool would be implemented. And secondly, it had to be extractable from the software models.

In terms of the user data availability, the total length of pipeline per diameter would not yet be known by those implementing the infrastructure estimation tool. Since infiltration volume is dependent on the length of pipeline per diameter, then the flow definition had to exclude infiltration flow. Similarly, since stormwater ingress is specified as a percentage spare capacity and the neither the infiltration volume nor pipe capacities would be known, stormwater ingress had to be excluded from the flow definition. Therefore, the only flow portion that could be accounted for was the user-generated or 'regular' flow during dry weather flow conditions. This could still be defined using ADDWF, PDDWF, or IPDWF.

It was considered that the IPDWF excluding infiltration was closest to the design flow and would therefore be the most effective flow definition, particularly for determining the pipe diameter distributions. However, this did not satisfy the second flow definition requirement, since the software did not make provision for this value to be accessed. As noted in Section 2.2.4.3, the software applied a lag time when routing the inflow hydrographs through the network, which dampened the combined peak or IPDWF. Therefore, the lag time was a function of the pipe lengths and diameters, but the model users would only be able to estimate IPDWF by directly summing the contributing inflow hydrographs. Therefore, the IPDWF for developing the infrastructure estimation tool would not be the same as the IPDWF estimated by the users of the tool, which would introduce an inherent error. The next-best option was then the PDDWF excluding infiltration, and this was readily accessible from the source data models.

D.3.2 Flow calculation

Calculating the PDDWF for each sample network was a straightforward process. The software provided a summary of the total daily flow volume generated per land use in terms of the 'Domestic' (or sewage return flow) and 'Leakage' flow. These flow components were calculated using the AADD method, where the AADD was obtained from billing data. The 'Domestic' and 'Leakage' components were summed to obtain the PDDWF flow contribution per land use in kilolitres per day. This flow breakdown per land use was used in determining the land use category of the sample network. The PDDWF contribution was summed for all land uses to obtain the total PDDWF for the sample network in kilolitres per day.

D.4 Land Use Grouping and Classification

D.4.1 Land use grouping

The source data models contained 14 unique land use types. In reality, service zones rarely comprise one single land use but rather a combination of land uses. From the flow calculations, the total PDDWF per land use was available. This was re-expressed as a percentage contribution to the total flow by each land use, effectively providing a land use breakdown for the service zone. Several options were available when incorporating this land use into the analysis. It could either be represented quantitatively as a land use distribution, or qualitatively in terms of the dominant land use in the service zone. The latter was selected as the preferred option since it would allow for a simpler analysis while still effectively describing the land use (provided the level of dominance was sufficient).

Representing land use in terms of one dominant land use made it a qualitative variable. In regression analysis, there are two ways of accounting for qualitative variables in models. One option is to model it as a categorical variable, where the categories are represented by integers that become independent variables in the model (Montgomery & Runger, 2014). The other option is to create separate regression models for each category. The latter was selected as the preferred option since it would be superior for modelling the category-specific effect of each independent variable.

However, obtaining sufficient sample networks to develop unique models for 14 different land uses was not practically achievable. This was because as service zones become larger, the land use becomes more mixed, and any dominance by one single land use is diminished. The land uses therefore needed to be combined logically. Three requirements were considered when grouping the land uses:

- Land uses that normally occur together should be grouped together.
- Land uses resulting in similar pipe layouts and spacing should be grouped together.
- The optimal number of groups would be the minimum such that within the groups the land uses were similar enough, and between the groups there was enough distinction.

The resulting land use groupings are presented in Table D-2. Table D-2 does not provide an exhaustive list of land uses, rather the ones that were present in the source data models. Therefore it is recommended that, when implementing the infrastructure estimation tool, the user should assign any land use not listed in Table D-2 to the land use category to which it most logically belongs.

Table D-2: Land use categories.

Land Use Category	Land Uses
General Residential	Very high income/ low density residential High income/ medium density residential Medium income/ high density residential Cluster Flats Farm/ agricultural holdings
Low Income Residential	Low income/ very high density residential
Non-Residential	Business/ commercial Educational Government/ institutional Industrial Mixed
Large	Large Public open space

D.4.2 Land use classification

A simple method was then followed to decide which land use category from Table D-2 best described each sample network. The percentage flow contributions per individual land use were summed for each land use category, and the category with the highest percentage contribution to the total flow for the sample network was the dominant land use for the network. It was important that the level of dominance was significant, because if the land uses were too mixed, then the effect of land use was at risk of being lost altogether. In 83% of samples, the dominant land use category contributed at least 65% of the flow, and in most of those cases the contribution was much higher. This was considered an acceptable level of dominance. This was also largely influenced by the area size, in that larger service zones tended to become more mixed with an associated decrease in the dominant land use's flow contribution percentage.

D.5 Real Surface Area Calculation

The XYZ coordinates of all DEM points lying within the service zone polygon were available for each network. The real surface area was then calculated by exporting these coordinates to MATLAB. The built-in 'Triangulation' function was used to connect a grid of triangles between the points, and then the 3D surface areas of the individual triangles were calculated and summed. Figure D-4 shows an example of the triangular grid from a top view. The irregular triangles along the edges in Figure D-4 did not fall within the service zone and were not included in the calculation. Each sample service zone's DEM was also rendered as a 3D surface plot and inspected for any irregularities in the DEM that would skew the topography factors. Figure D-5 shows the surface plot of an acceptable network. Figure D-6 shows the plot of a network with an irregularity in the DEM, which was consequently removed from the dataset.

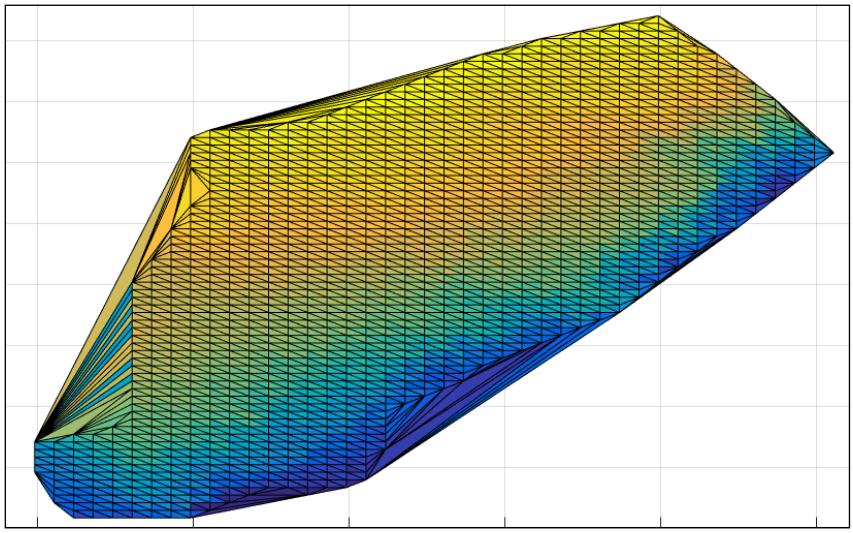


Figure D-4: Connection of triangles between DEM points in a sample service zone.

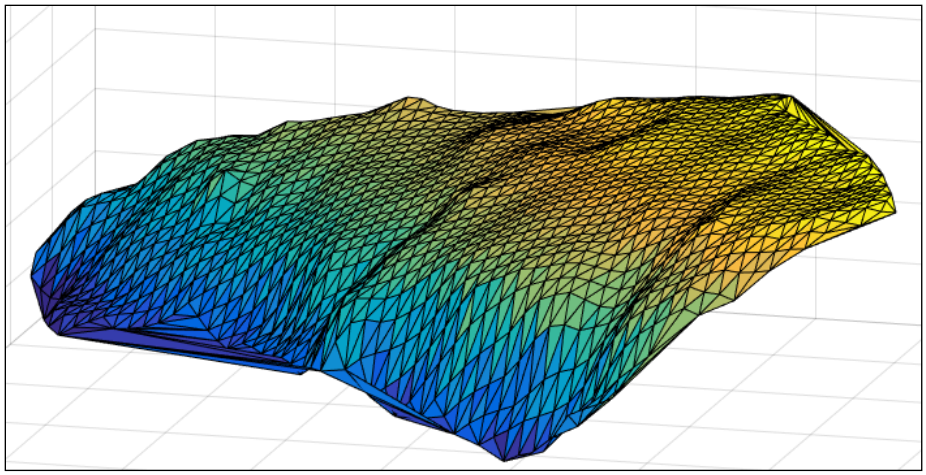


Figure D-5: Acceptable 3D surface plot of a sample service zone.

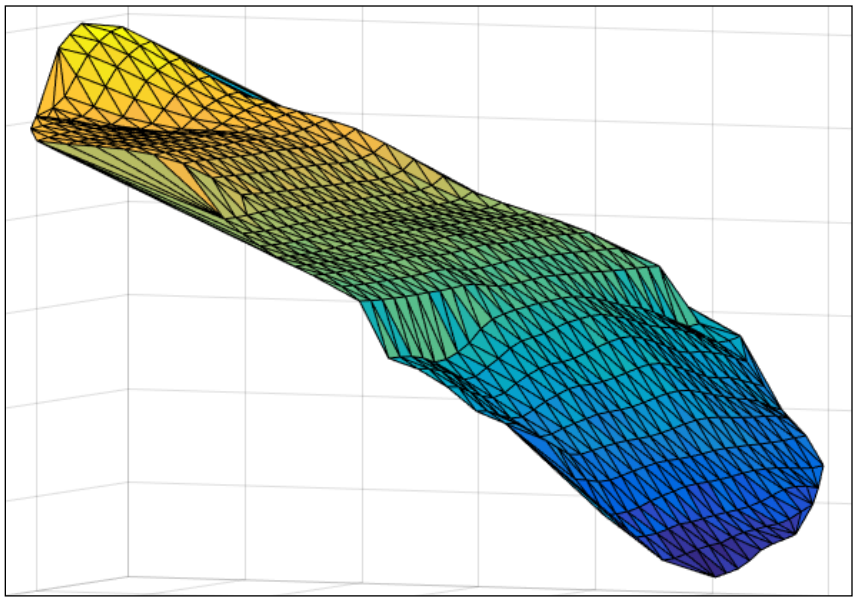


Figure D-6: Unacceptable 3D surface plot of a sample service zone.

Appendix E
STUDY OUTCOME I: MODEL DEVELOPMENT RESULTS

E.1 Results from Step 2a: Standard Backward Elimination

Table E-1 (continued on the next page) provides the results from the standard backward elimination process in Step 2a. The 'Variables' rows indicate the variables used in each model. For each model, the 'x's on the left indicate the starting combination of variables, and the '**x**'s on the right indicate the end combination of variables after the insignificant variables ($p > 0.05$) were removed. The full model formulae are not shown as they were not of interest at this stage. It is noted that the OLS assumptions were checked before any of the results were considered reliable.

Table E-1: Model results from Step 2a: Standard backward elimination.

	Regression Model Number		2a-A	2a-B	2a-C	2a-D	2a-E	2a-F	2a-G	2a-H *	2a-I
Variables	Y	Total pipeline length	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_1	Plane area	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_2	PDDWF per hectare UHs per hectare	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_3	Circularity ratio	x **x**	x **x**	x	x	x	x	x	x	x
	X_4	Centroid-mouth relative radius	x	x	x	x	x	x	x	x	x
	X_5	Mean perimeter slope	x **x**								
		Mean basin slope		x **x**							
		Melton's ruggedness			x **x**						
		Surface area ratio				x **x**					
		Total relief					x **x**				
		Mean relief						x **x**			
		Elevation SD							x **x**		
		Ruggedness number								x **x**	
		Deviation from mean elevation									x **x**
Results	R^2		0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.95
	Adjusted R^2		0.95	0.95	0.95	0.95	0.96	0.96	0.95	0.96	0.95
	R^2 test data		0.94	0.94	0.95	0.95	0.95	0.96	0.95	0.96	0.95
	Log-likelihood		-419	-420	-422	-423	-417	-410	-418	-400	-424
	AIC		848	849	851	855	841	828	843	808	857
	BIC		865	866	864	867	854	841	856	821	870
	MAE test data (km)		1.58	1.60	1.51	1.49	1.49	1.43	1.44	1.37	1.49
	RMSE test data (km)		2.41	2.43	2.31	2.33	2.29	2.13	2.26	2.08	2.27

Table E-1 (continued).

	Regression Model Number	2a-J	2a-K	2a-L	2a-M	2a-N	2a-O	2a-P	2a-Q *	2a-R	2a-S	2a-T
Variables	Y Total pipeline length	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_1 Plane area	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_2 PDDWF per hectare										x **x**	
	X_2 UHs per hectare	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**		
	X_3 Circularity ratio	x	x **x**	x	x	x	x	x	x	x		
	X_4 Centroid-mouth relative radius	x	x	x	x	x	x	x	x	x		
	X_5 Mean perimeter slope	x **x**										
	X_5 Mean basin slope		x **x**									
	X_5 Melton's ruggedness			x **x**								
	X_5 Surface area ratio				x **x**							
	X_5 Total relief					x **x**						
	X_5 Mean relief						x **x**					
	X_5 Elevation SD							x **x**				
	X_5 Ruggedness number								x **x**			
	X_5 Deviation from mean elevation									x		
Results	R^2	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.95	0.95	0.95
	Adjusted R^2	0.95	0.95	0.95	0.95	0.96	0.96	0.95	0.96	0.95	0.95	0.95
	R^2 test data	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.95
	Log-likelihood	-422	-421	-422	-424	-417	-411	-418	-399	-428	-427	-433
	AIC	852	852	852	857	842	829	844	806	862	859	870
	BIC	865	868	865	870	854	842	857	819	872	869	876
	MAE test data (km)	1.35	1.47	1.36	1.38	1.35	1.27	1.29	1.26	1.42	1.47	1.41
	RMSE test data (km)	2.23	2.34	2.25	2.28	2.23	2.05	2.20	2.04	2.28	2.32	2.31

* Models 2a-H and 2a-Q were disqualified as valid models based on an additional check that proved the ruggedness number to be an unreliable variable (see Appendix E.2)

E.2 Ruggedness Number Check in Step 2a: Standard Backward Elimination

The candidate independent variable, ruggedness number, was a special-case variable due to the nature of its definition. Therefore, it had to undergo a check as to its reliability. The formula for the ruggedness number as previously defined in Chapter 4 is presented in Equation E-1.

$$Ruggedness\ Number = \frac{Total\ Pipeline\ Length \times (H_{max} - H_{min})}{1000 \times Area} \quad \text{E-1}$$

In this study, $H_{max} - H_{min}$ was defined as the variable named total relief, and area was represented by the variable named plane area. Equation E-1 can then be written as in Equation E-2:

$$Ruggedness\ Number = \frac{Total\ Pipeline\ Length \times Total\ Relief}{1000 \times Plane\ Area} \quad \text{E-2}$$

When quantifying the sample network characteristics to obtain the dataset, the ruggedness number was calculated using Equation E-2. Therefore, the dependent variable, total pipeline length, was used as a direct input for the independent variable, ruggedness number. The ruggedness number thus calculated was used for the model development in Step 2a, and for quantifying the associated model performance. Consequently, the two models (2a-H and 2a-Q) that included the ruggedness number as an independent variable had the best performance results in Step 2a.

In reality, however, the total pipeline length would be an unknown factor. Therefore, a regression model using ruggedness number as an independent variable would have to be rearranged so that the total pipeline length occurred only on the left hand side (LHS) of the model equation. With the total pipeline length no longer incorporated in this independent variable, an associated decrease in the model prediction accuracy would be expected; therefore a check was required to ascertain whether the decrease was acceptable.

Only one of the ruggedness number models, 2a-Q, was used for the check. The check was done by comparing the performance results of three regression models, namely: model 2a-Q; model 2a-Q rearranged such that the total pipeline length was on the LHS only; and model 2a-F, which was the next-best model in terms of the log-likelihood, AIC and BIC. The formula for regression model 2a-Q is defined in Equation E-3:

$$Total\ Pipeline\ Length = -1.567 + 0.121(Plane\ Area) + 0.117(UHs\ Per\ Hectare) + 0.259(Ruggedness\ Number) \quad \text{E-3}$$

Substituting Equation E-2 for the ruggedness number in Equation E-3 yields Equation E-4. Then, rearranging Equation E-4 so that total pipeline length is on the LHS only, yields Equation E-5, which represents the rearranged model 2a-Q.

$$Total\ Pipeline\ Length = -1.567 + 0.121(Plane\ Area) + 0.117(UHs\ Per\ Hectare) + 0.259\left(\frac{Total\ Pipeline\ Length \times Total\ Relief}{1000 \times Plane\ Area}\right) \quad \text{E-4}$$

$$Total\ Pipeline\ Length = \frac{-1.567 + 0.121(Plane\ Area) + 0.117(UHs\ Per\ Hectare)}{1 - \frac{0.259(Total\ Relief)}{1000(Plane\ Area)}} \quad \text{E-5}$$

The formula for model 2a-F, the next-best model from Step 2a that did not use ruggedness number, is defined in Equation E-6:

$$Total\ Pipeline\ Length = -1.532 + 0.119(Plane\ Area) + 0.133(UHs\ Per\ Hectare) + 0.081(Mean\ Relief) \quad \text{E-6}$$

The three models defined by Equations E-4, E-5 and E-6 were applied to the General Residential test dataset. The set of residuals (difference between the actual and predicted values) was used to calculate the test data R^2 and the average error in terms of the mean absolute error (MAE) and root mean squared error (RMSE). The results are presented in Table E-2.

Table E-2: Performance results from the ruggedness number check.

Regression Model	2a-Q	Rearranged 2a-Q	2a-F
Equation	Equation E-4	Equation E-5	Equation E-6
R^2 Test Data	0.96	0.92	0.96
MAE (km)	1.26	1.94	1.43
RMSE (km)	2.04	2.90	2.13

As expected, the rearranged 2a-Q had a lower test data R^2 and a higher average error than the original 2a-Q. Notably, the MAE was on average nearly 700 m less accurate, and the RMSE was on average over 800 m less accurate. Moreover, model 2a-F had a higher test data R^2 and a lower average error than the rearranged 2a-Q. Since the ruggedness number results were unreliable, and since a stronger model that did not use the ruggedness number existed, the ruggedness number was removed from the set of candidate variables and excluded from any further analyses.

E.3 Results from Step 2b: Principal Component Analysis (PCA)

Table E-3 provides the results from the PCA models in Step 2b. The 'Variables in Principal Components' rows indicate the variables used to generate the principal components, and the 'Number of Significant Principal Components' row indicates the number of principal components remaining in each model after those with $p > 0.05$ had been removed. The make-up of the principal components and the full model formulae are not shown as they were not of interest at this stage. It is noted that the OLS assumptions were checked before any of the results were considered reliable.

Table E-3: Model results from Step 2b: Principal component analysis.

Regression Model Number			**2b-A**	**2b-B**	**2b-C**
Variables in Principal Components	**Y**	Total pipeline length	x	x	x
	X_1	Plane area	x	x	x
	X_2	PDDWF per hectare	x	x	x
		UHs per hectare	x	x	x
	X_5	Mean perimeter slope	x	x	
		Mean basin slope	x		
		Melton's ruggedness	x	x	
		Surface area ratio	x		
		Total relief	x	x	
		Mean relief	x		x
		Elevation SD	x		
Number of Significant Principal Components			7	3	3
Results	R^2		0.96	0.96	0.96
	Adjusted R^2		0.96	0.96	0.96
	R^2 test data		0.96	0.95	0.96
	Log-likelihood		-406	-415	-409
	AIC		828	839	827
	BIC		854	852	840
	MAE test data (km)		1.34	1.41	1.35
	RMSE test data (km)		2.05	2.25	2.09

E.4 Results from Step 3a: Weighted Least Squares Regression (WLS)

Table E-4 provides the results from the WLS regression models compiled in Step 3a. The 'Variables' rows indicate the variables used in each model. For each model, the 'x's on the left indicate the starting combination of variables, and the '**x**'s on the right indicate the end combination of variables after the insignificant variables ($p > 0.05$) were removed. The full model formulae are not shown as they were not of interest at this stage. It is noted that the OLS assumptions were checked before any of the results were considered reliable.

Table E-4: Model results from Step 3a: Weighted least squares regression.

Regression Model Number			**2-A**		**2-B**		**2-C**		**2-D**		**2-E**		**2-F**	
Weighting Method			Weights 1				Weights 2				Weights 3			
Variables	Y	Total pipeline length	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**
	X_1	Plane area	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**
	X_2	PDDWF per hectare	x	**x**			x	**x**			x	**x**		
		Number of UH's per hectare			x				x	**x**			x	**x**
	X_5	Mean relief	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**
Results	R^2		0.94		0.95		0.92		0.92		0.94		0.96	
	Adjusted R^2		0.94		0.95		0.92		0.91		0.94		0.95	
	R^2 test data		0.96		0.96		0.96		0.96		0.96		0.96	
	Log-likelihood		-343		-333		-345		-350		-327		-321	
	AIC		694		675		698		708		662		650	
	BIC		707		688		711		721		675		663	
	MAE test data (km)		1.27		1.20		1.44		1.42		1.33		1.27	
	RMSE test data (km)		2.04		2.07		2.12		2.11		2.08		2.03	

E.5 Results from Step 4: Checking Variable Conclusions

Table E-5 (continued on the next page) provides the results from the standard backward elimination process in Step 4. The 'Variables' rows indicate the variables used in each model. For each model, the 'x's on the left indicate the starting combination of variables, and the '**x**'s on the right indicate the end combination of variables after the insignificant variables ($p > 0.05$) were removed. The full model formulae are not shown as they were not of interest at this stage. It is noted that the OLS assumptions were checked before any of the results were considered reliable.

Table E-5: Model results from Step 4: Checking variable conclusions.

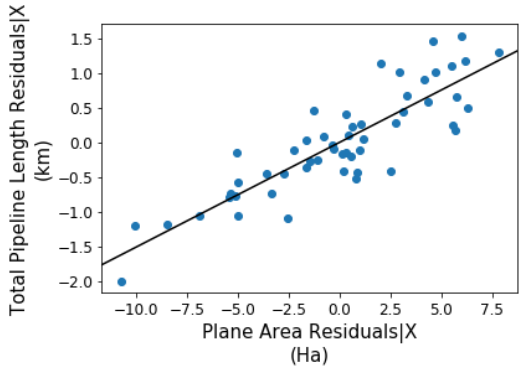
Regression Model Number			**4-A**		**4-B**		**4-C**		**4-D**		**4-E**		**4-F**		**4-G**		**4-H**	
Weighting Method			Weights 1															
Variables	Y	Total pipeline length	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**
	X_1	Plane area	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**
	X_2	PDDWF per hectare UHs per hectare	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**	x	**x**
	X_3	Circularity ratio	x		x		x		x		x		x		x		x	
	X_4	Centroid-mouth relative radius	x		x		x		x		x		x		x		x	
	X_5	Mean perimeter slope	x															
		Mean basin slope			x													
		Melton's ruggedness					x											
		Surface area ratio							x									
		Total relief									x	**x**						
		Mean relief											x	**x**				
		Elevation standard deviation													x	**x**		
		Deviation from mean elevation															x	**x**
Results	R^2		0.96								0.96		0.96		0.96		0.96	
	Adjusted R^2		0.96								0.96		0.96		0.96		0.96	
	R^2 test data		0.95								0.95		0.96		0.95		0.94	
	Log-likelihood		-156								-151		-147		-152		-150	
	AIC		317								311		303		312		308	
	BIC		325								321		313		322		318	
	MAE test data (km)		1.53								1.50		1.47		1.50		1.58	
	RMSE test data (km)		2.64								2.47		2.43		2.47		2.71	

Table E-5 (continued).

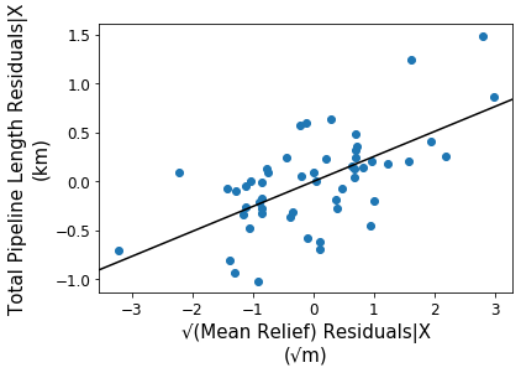
	Regression Number	4-I	4-J	4-K	4-L	4-M	4-N	4-O	4-P	4-Q	4-R
	Weighting Method	Weights 1									
Variables	Y Total pipeline length	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_1 Plane area	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**
	X_2 PDDWF per hectare										
	X_2 UHs per hectare	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	x **x**	
	X_3 Circularity ratio	x	x	x	x	x	x	x	x		
	X_4 Centroid-mouth relative radius	x **x**	x **x**	x **x**	x **x**	x **x**	x	x	x **x**		
	X_5 Mean perimeter slope	x									
	X_5 Mean basin slope		x								
	X_5 Melton's ruggedness			x							
	X_5 Surface area ratio				x						
	X_5 Total relief					x **x**					
	X_5 Mean relief						x **x**				
	X_5 Elevation standard deviation							x **x**			
	X_5 Dev. from mean elevation								x **x**		
Results	R^2	0.96				0.96	0.96	0.96	0.96	0.96	0.95
	Adjusted R^2	0.96				0.96	0.96	0.96	0.96	0.96	0.95
	R^2 test data	0.95				0.96	0.96	0.96	0.94	0.95	0.94
	Log-likelihood	-153				-147	-146	-150	-149	-156	-164
	AIC	314				304	300	308	308	317	332
	BIC	324				316	310	318	320	325	337
	MAE test data (km)	1.54				1.49	1.47	1.48	1.61	1.51	1.54
	RMSE test data (km)	2.65				2.42	2.43	2.44	2.73	2.67	2.75

E.6 Partial Regression Plots for the Final Study Outcome I Models

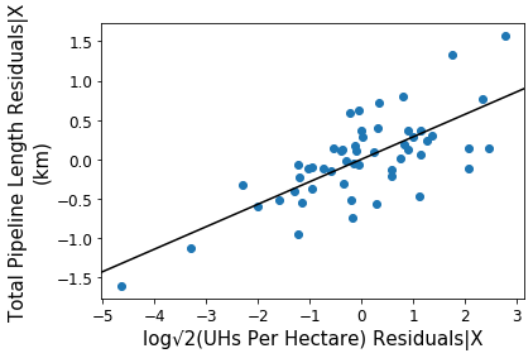
Figure E-1 to Figure E-9 present the partial regression plots for the final total pipeline length models. For each model category, the plots illustrate the relationship between the dependent variable and each independent variable in the model, after the influence of the other independent variables in the model has been accounted for. It is noted that some of the models did not use all three available independent variables (plane area, mean relief and UHs per hectare). The plots were used to visually assess the correlation strength, verify p-value conclusions, and to identify outliers and influential points.



(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare)

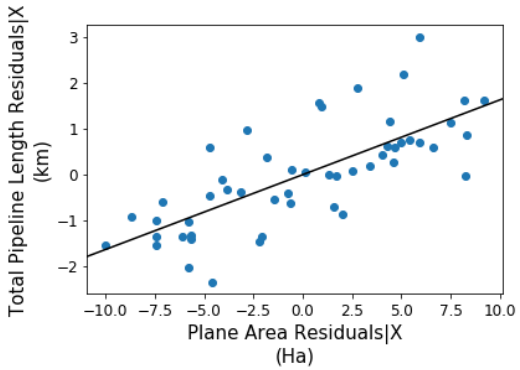


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area and $\log_{\sqrt{2}}$(UHs per hectare)

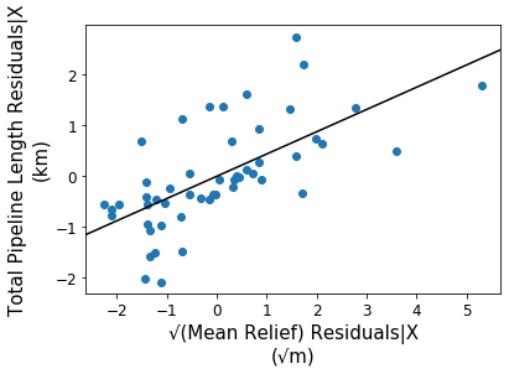


(iii) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare) residuals, controlling for plane area and $\sqrt{\text{mean relief}}$

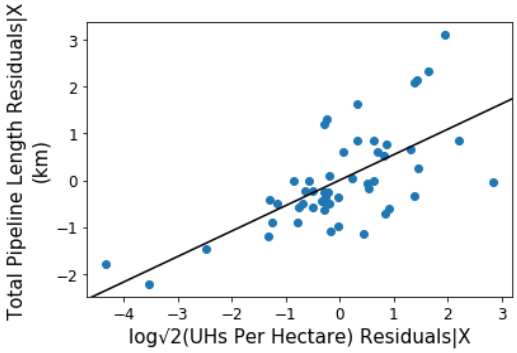
Figure E-1: Partial regression plots for model 'General Residential', 0 – 20 ha.



(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare)

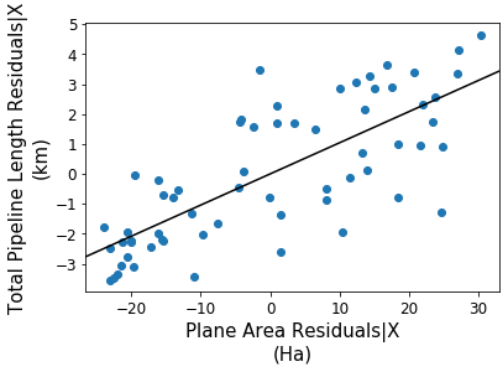


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area and $\log_{\sqrt{2}}$(UHs per hectare)

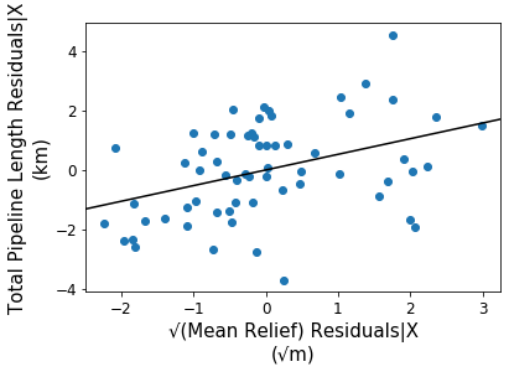


(iii) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare) residuals, controlling for plane area and $\sqrt{\text{mean relief}}$

Figure E-2: Partial regression plots for model 'General Residential', 20 – 40 ha.

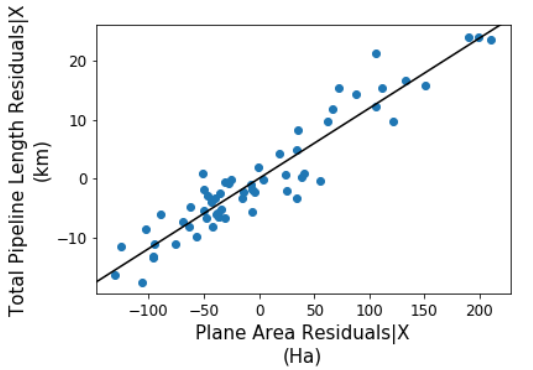


(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$

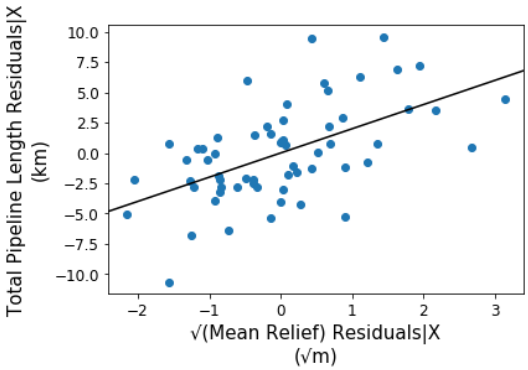


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area

Figure E-3: Partial regression plots for model 'General Residential', 40 – 100 ha.

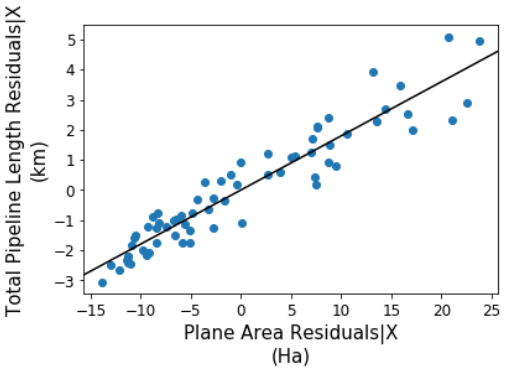


(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$

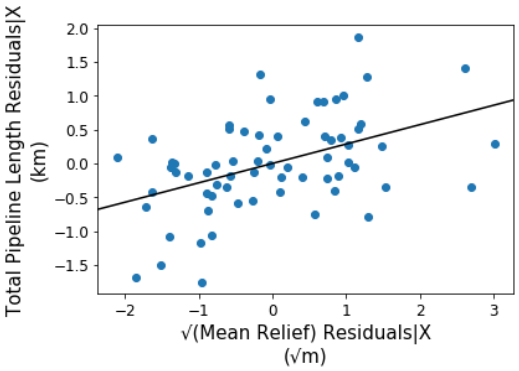


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area

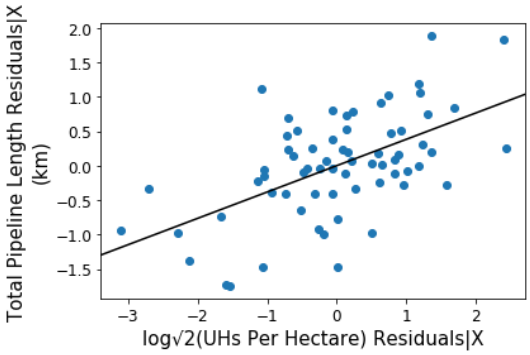
Figure E-4: Partial regression plots for model 'General Residential', 100 – 450 ha.



(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare)

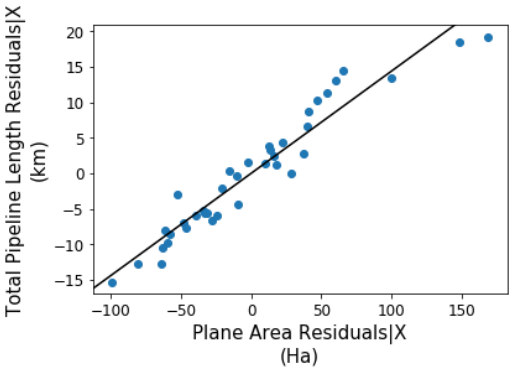


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area and $\log_{\sqrt{2}}$(UHs per hectare)

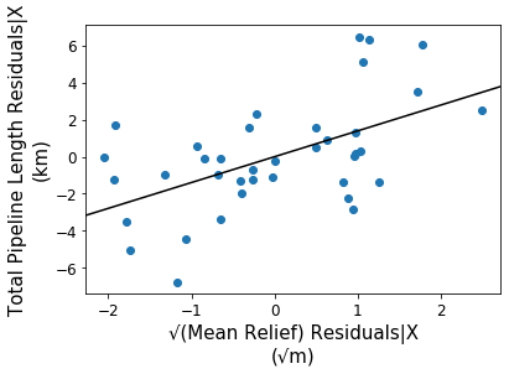


(iii) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare) residuals, controlling for plane area and $\sqrt{\text{mean relief}}$

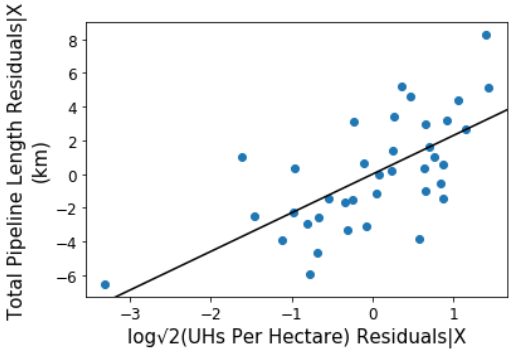
Figure E-5: Partial regression plots for model 'Low Income Residential', 0 – 40 ha.



(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare)

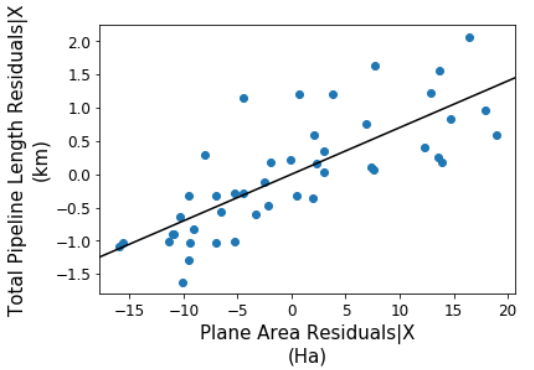


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area and $\log_{\sqrt{2}}$(UHs per hectare)

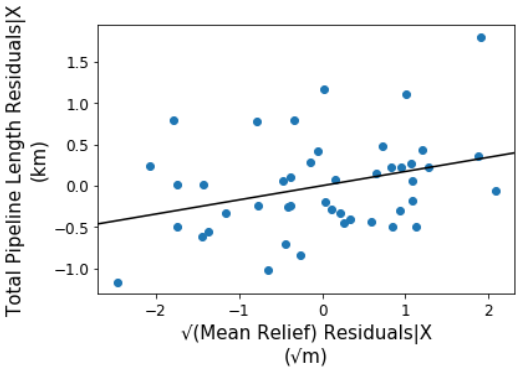


(iii) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare) residuals, controlling for plane area and $\sqrt{\text{mean relief}}$

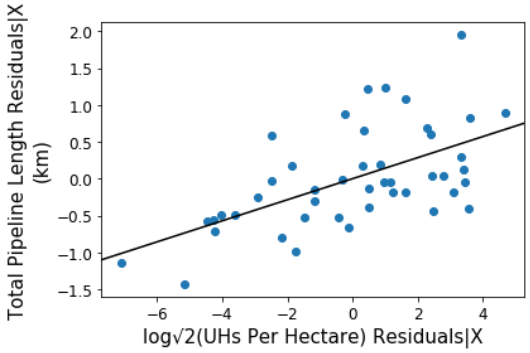
Figure E-6: Partial regression plots for model 'Low Income Residential', 40 – 300 ha.



(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$ and $\log_{\sqrt{2}}$(UHs per hectare)

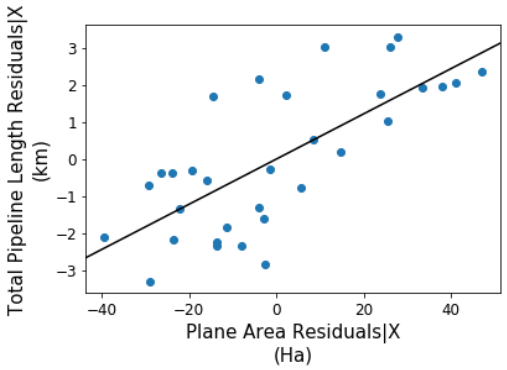


(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area and $\log_{\sqrt{2}}$(UHs per hectare)

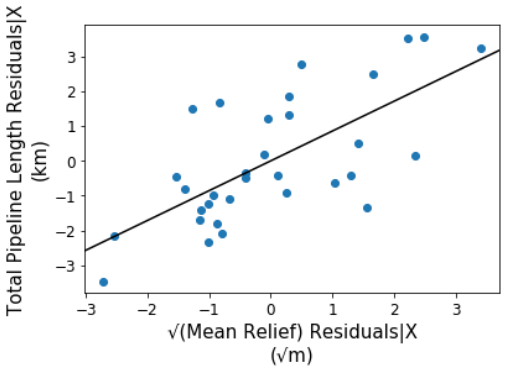


(iii) Total pipeline length residuals vs. $\log_{\sqrt{2}}$(UHs per hectare) residuals, controlling for plane area and $\sqrt{\text{mean relief}}$

Figure E-7: Partial regression plots for model 'Non-Residential', 0 – 40 ha.

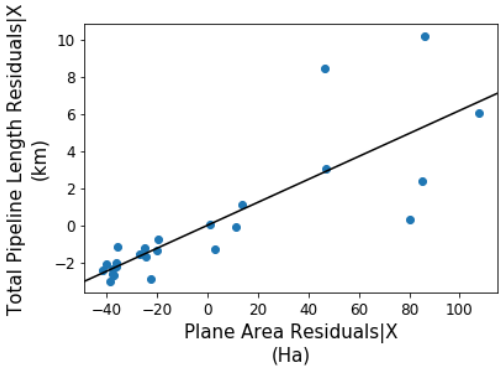


(i) Total pipeline length residuals vs. plane area residuals, controlling for $\sqrt{\text{mean relief}}$



(ii) Total pipeline length residuals vs. $\sqrt{\text{mean relief}}$ residuals, controlling for plane area

Figure E-8: Partial regression plots for model 'Non-Residential', 40 – 120 ha.



(i) Total pipeline length residuals vs. plane area residuals

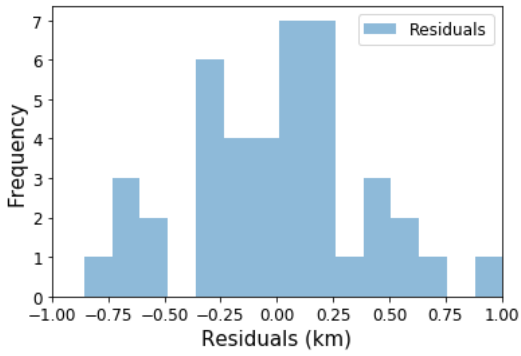
Figure E-9: Partial regression plots for model 'Large', 0 – 160 ha.

E.7 OLS Assumption Check Plots for the Final Study Outcome I Models

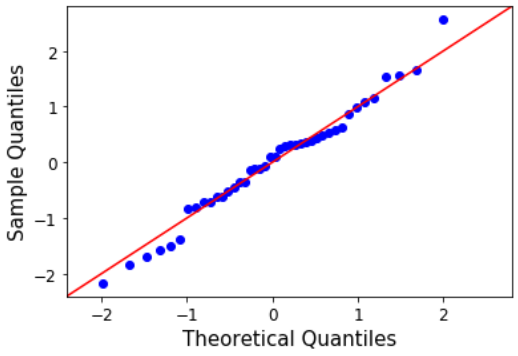
Figure E-10 to Figure E-18 present the plots which were used to verify that the OLS assumptions for the final nine total pipeline length models were satisfied. The following plots are displayed:

- Residuals histogram for normality.
- Normal probability plot for normality.
- Weighted residual plots versus the dependent and independent variables for linearity and adequate treatment of non-constant variance.
- Residual plots versus the observation order for independence.

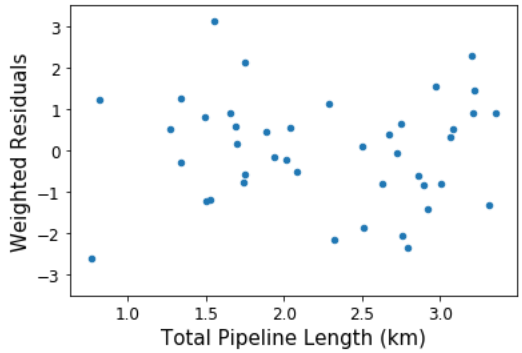
For each model, the plots indicated that the OLS assumptions were adequately satisfied.



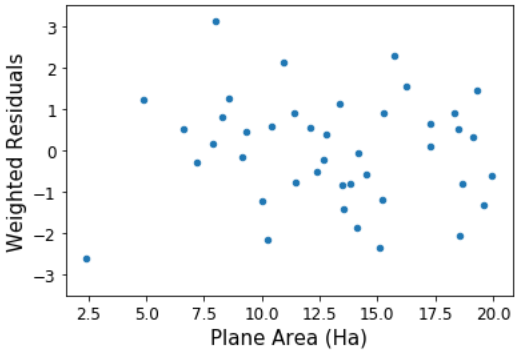
(i) Residuals histogram



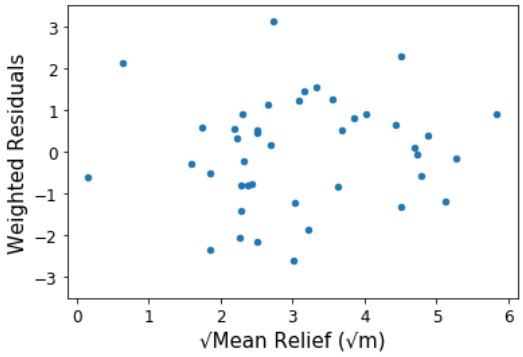
(ii) Normal probability plot



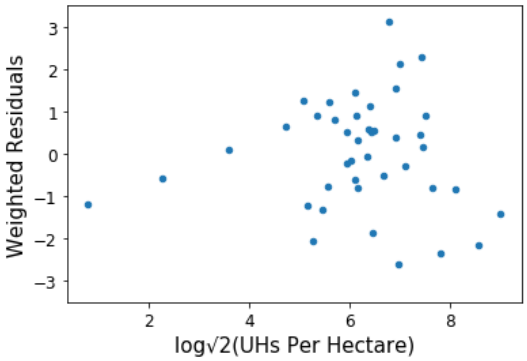
(iii) Weighted residuals vs. total pipeline length



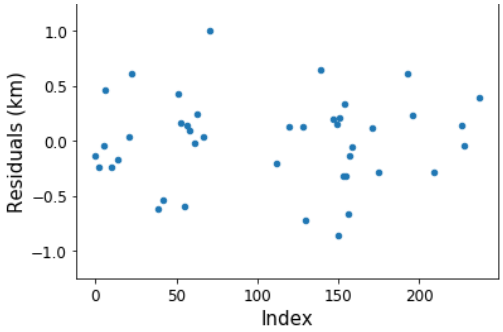
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

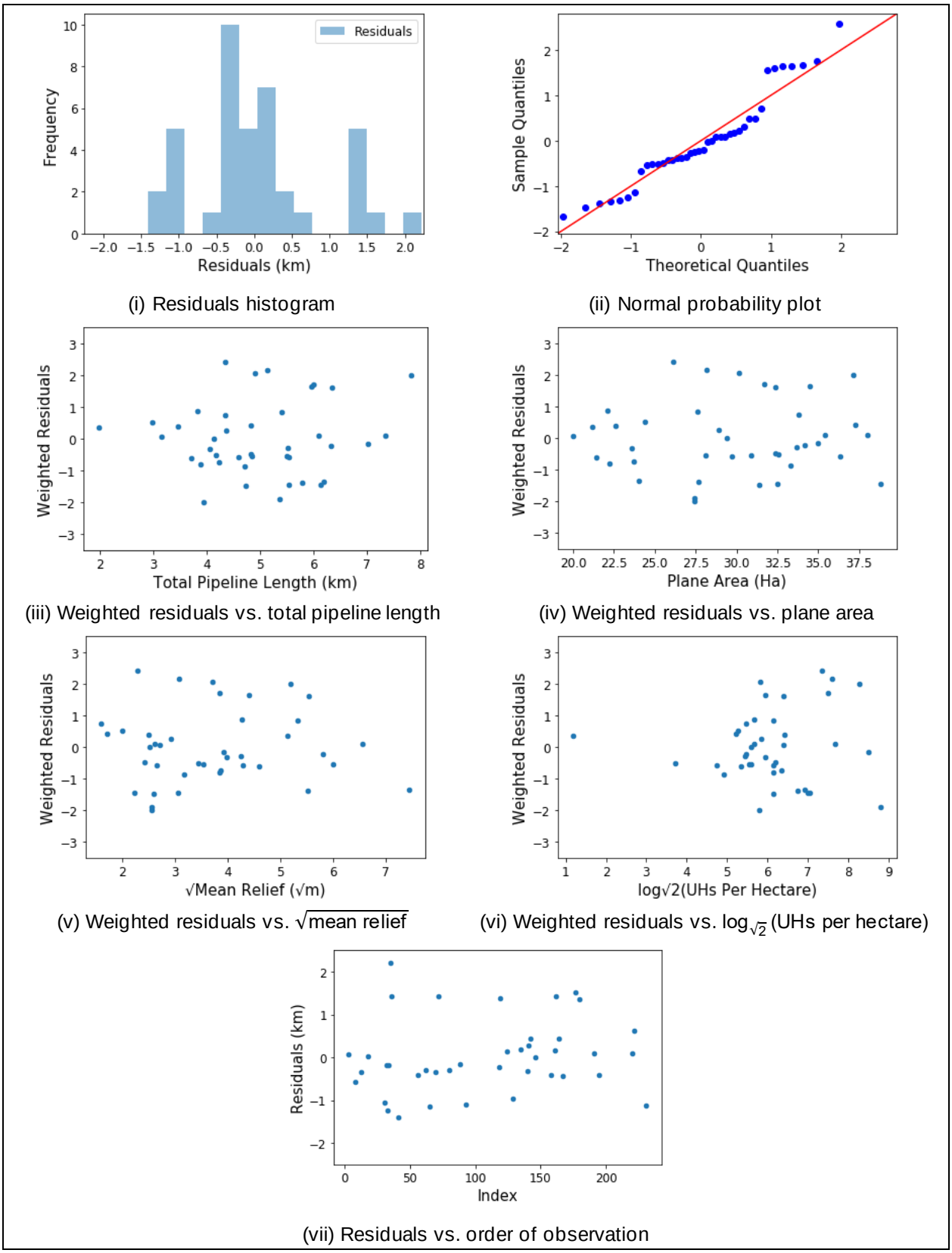


(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)



(vii) Residuals vs. order of observation

Figure E-10: OLS assumption check plots ('General Residential', 0 – 20 ha).



(i) Residuals histogram

(ii) Normal probability plot

(iii) Weighted residuals vs. total pipeline length

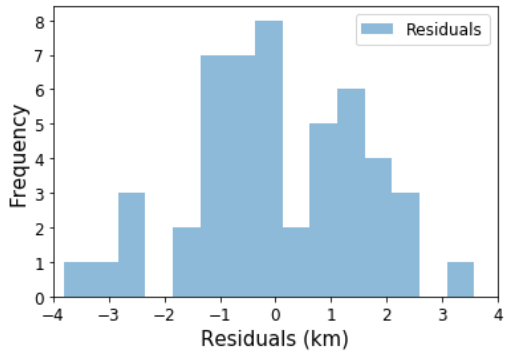
(iv) Weighted residuals vs. plane area

(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

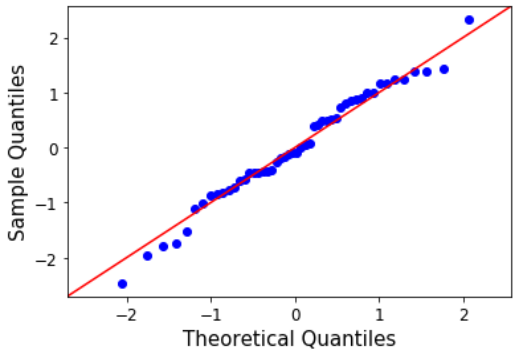
(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

(vii) Residuals vs. order of observation

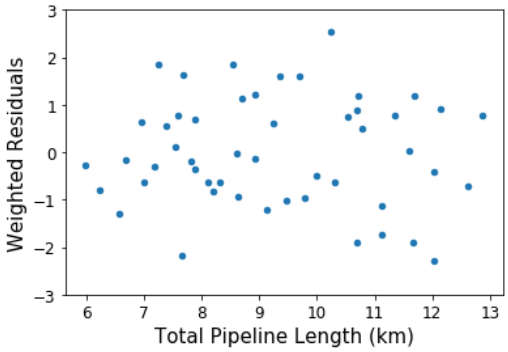
Figure E-11: OLS assumption check plots ('General Residential', 20 – 40 ha).



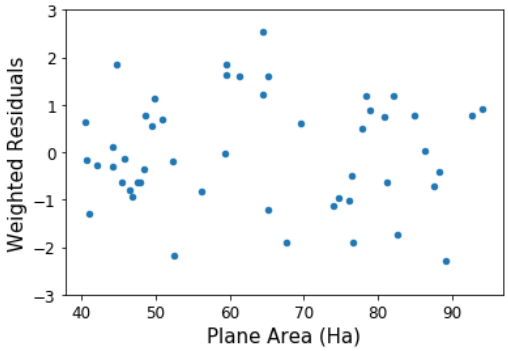
(i) Residuals histogram



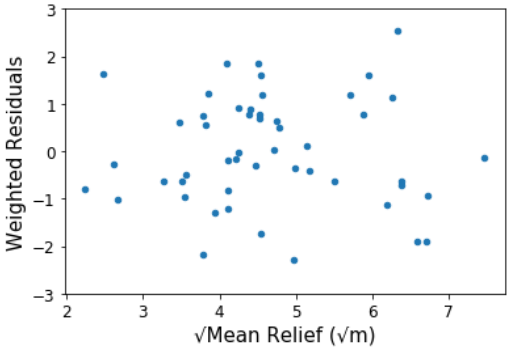
(ii) Normal probability plot



(iii) Weighted residuals vs. total pipeline length



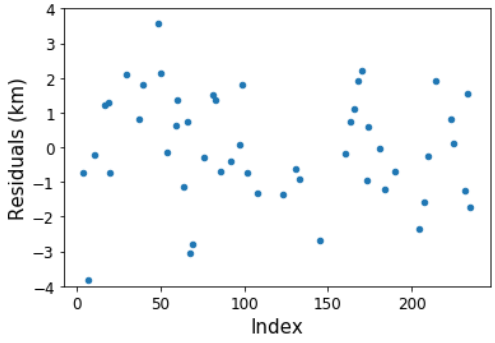
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

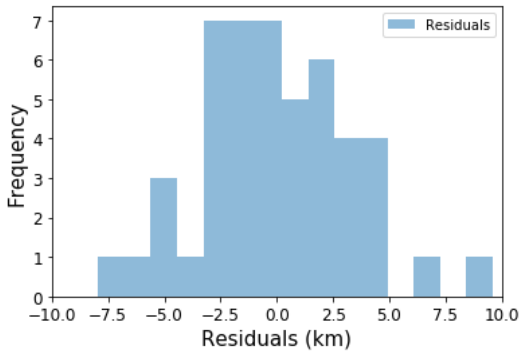
N/A

(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

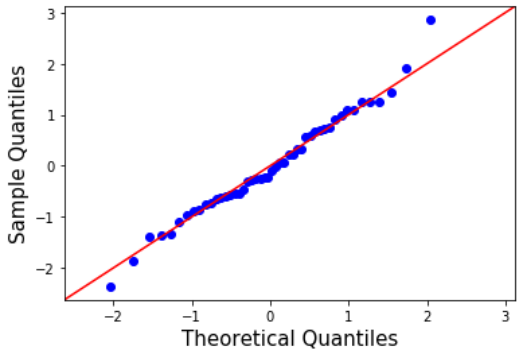


(vii) Residuals vs. order of observation

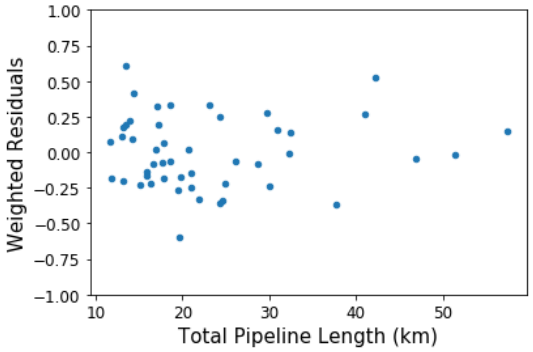
Figure E-12: OLS assumption check plots ('General Residential', 40 – 100 ha).



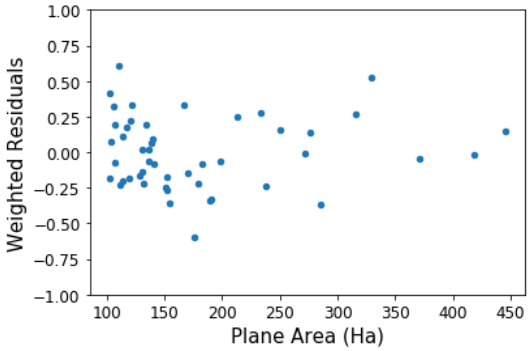
(i) Residuals histogram



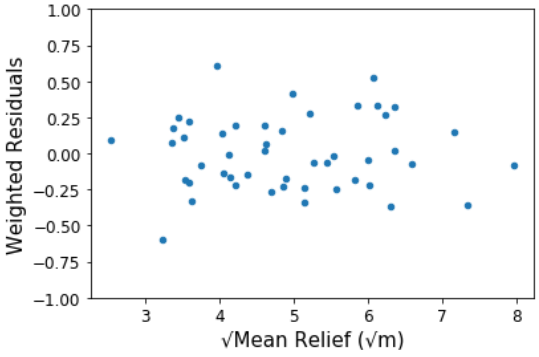
(ii) Normal probability plot



(iii) Weighted residuals vs. total pipeline length



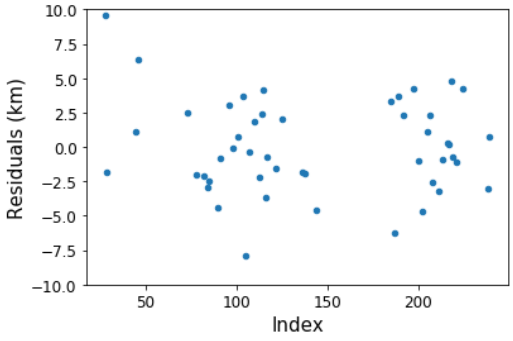
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

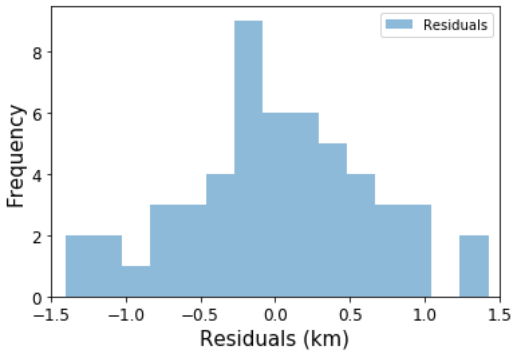
N/A

(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

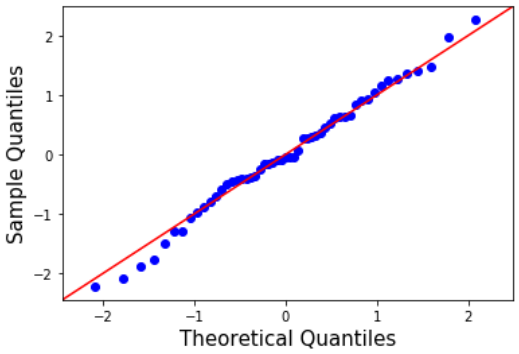


(vii) Residuals vs. order of observation

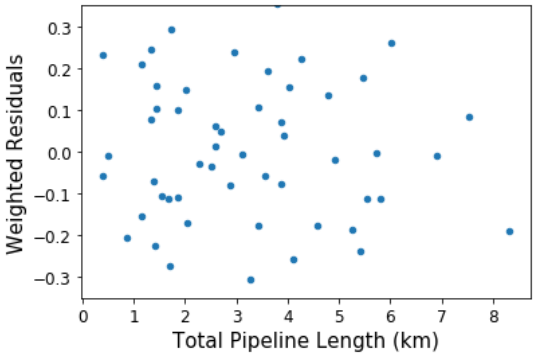
Figure E-13: OLS assumption check plots ('General Residential', 100 – 450 ha).



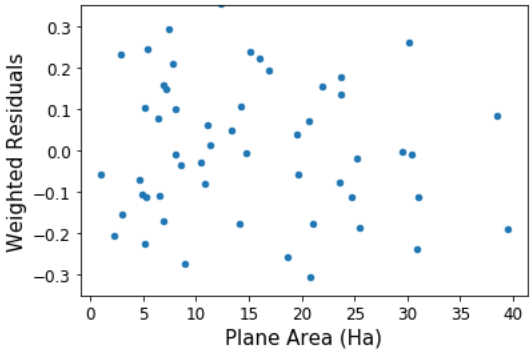
(i) Residuals histogram



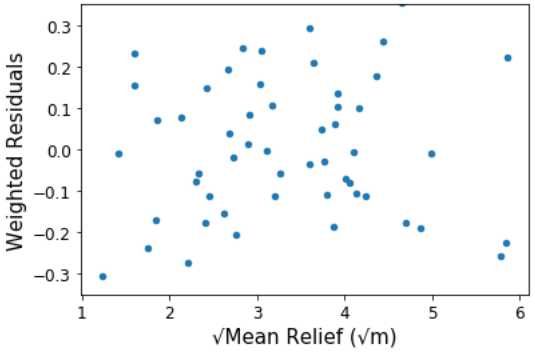
(ii) Normal probability plot



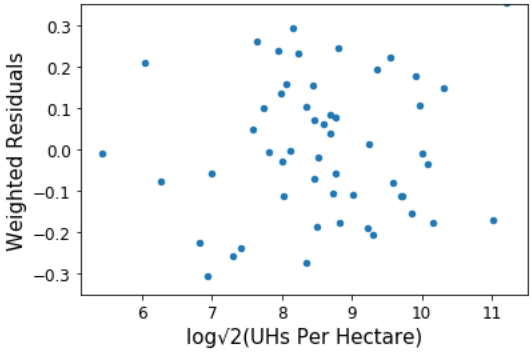
(iii) Weighted residuals vs. total pipeline length



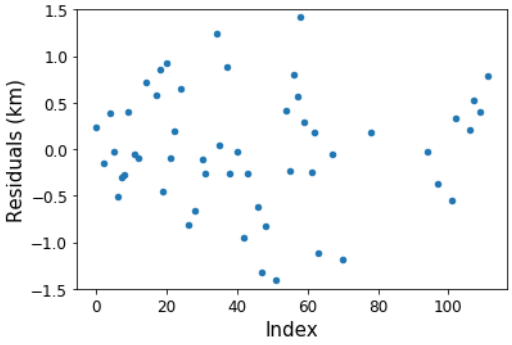
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

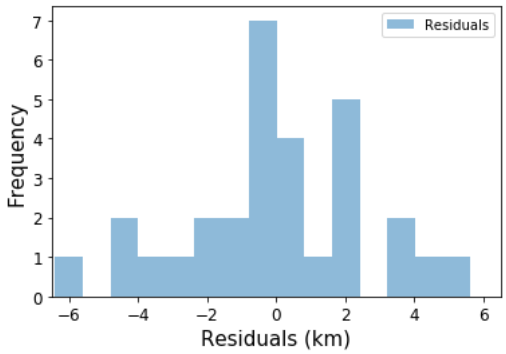


(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

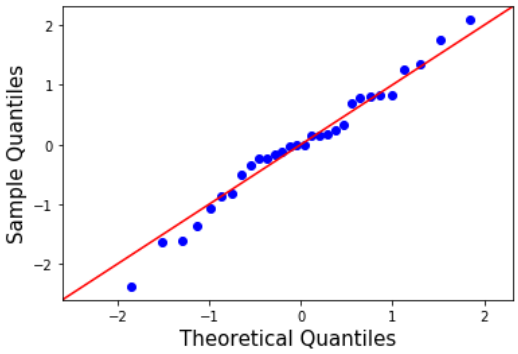


(vii) Residuals vs. order of observation

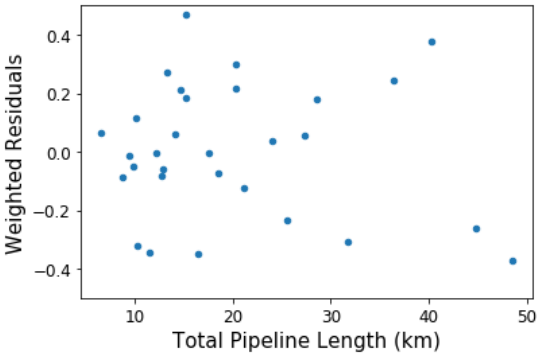
Figure E-14: OLS assumption check plots ('Low Income Residential', 0 – 40 ha).



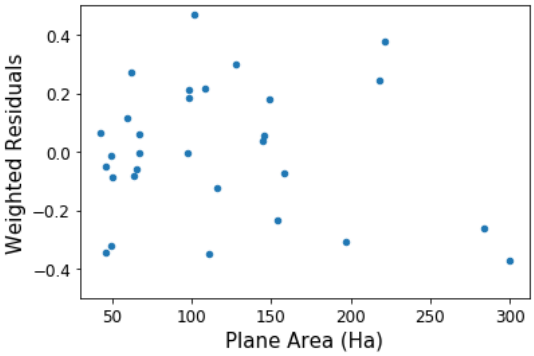
(i) Residuals histogram



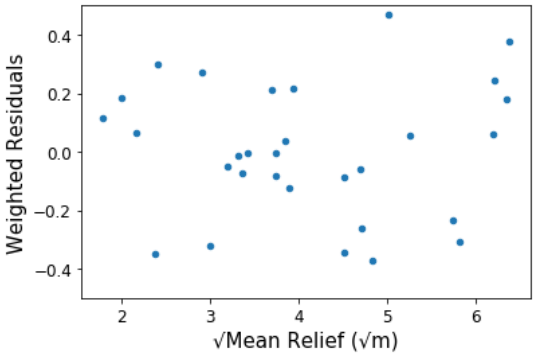
(ii) Normal probability plot



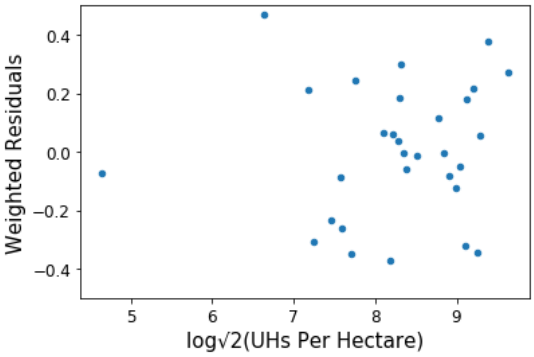
(iii) Weighted residuals vs. total pipeline length



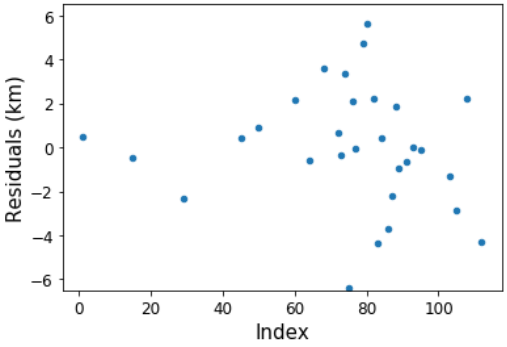
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

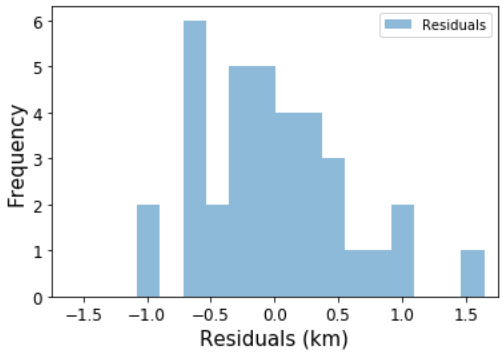


(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

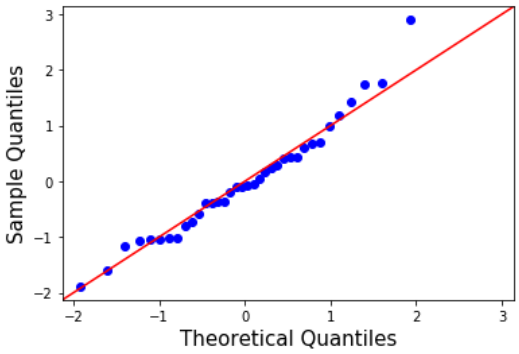


(vii) Residuals vs. order of observation

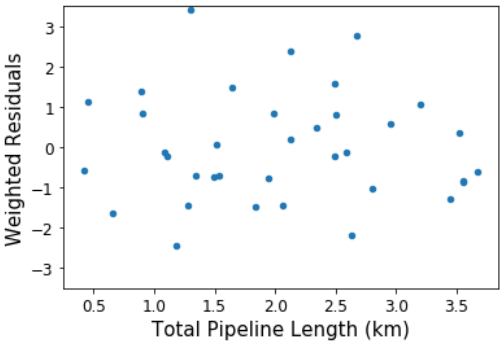
Figure E-15: OLS assumption check plots ('Low Income Residential', 40 – 300 ha).



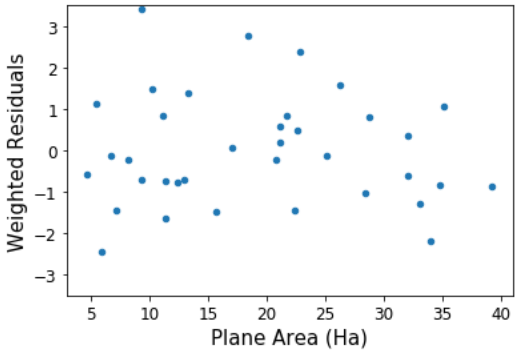
(i) Residuals histogram



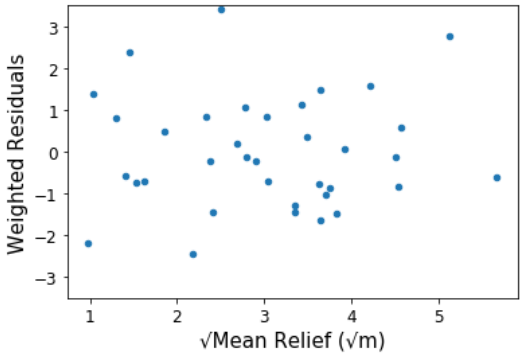
(ii) Normal probability plot



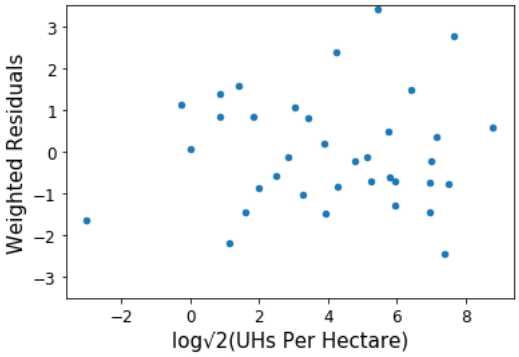
(iii) Weighted residuals vs. total pipeline length



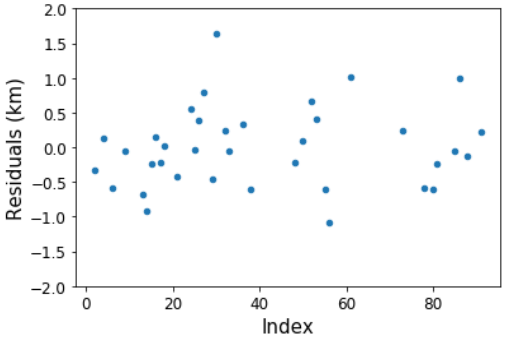
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

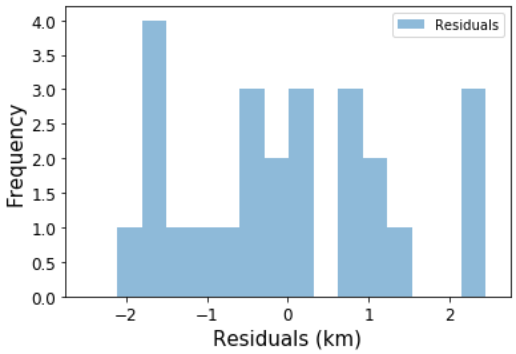


(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)

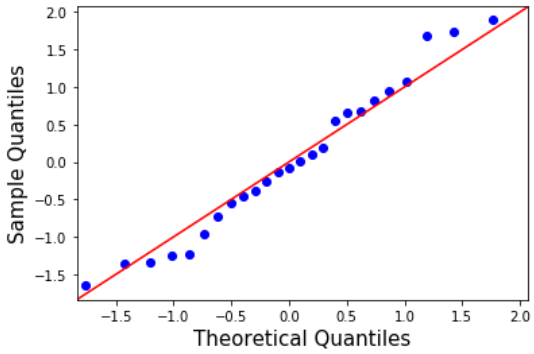


(vii) Residuals vs. order of observation

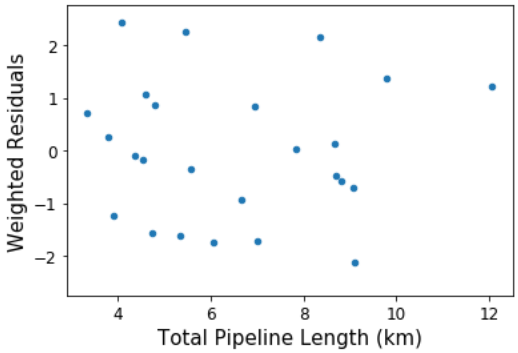
Figure E-16: OLS assumption check plots ('Non-Residential', 0 – 40 ha).



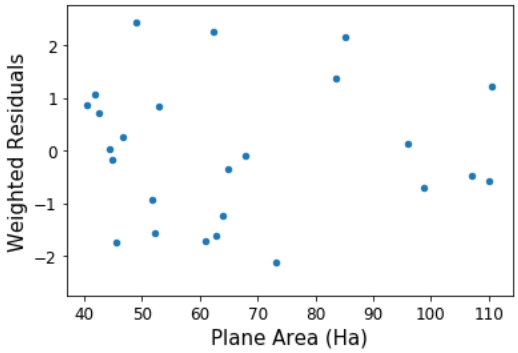
(i) Residuals histogram



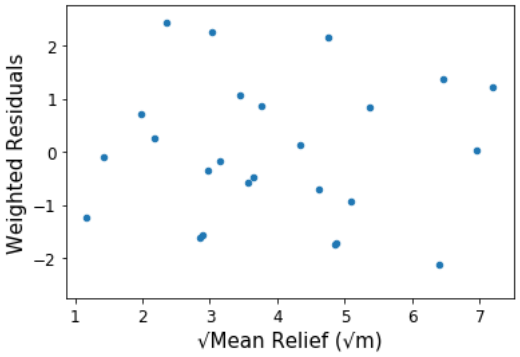
(ii) Normal probability plot



(iii) Weighted residuals vs. total pipeline length



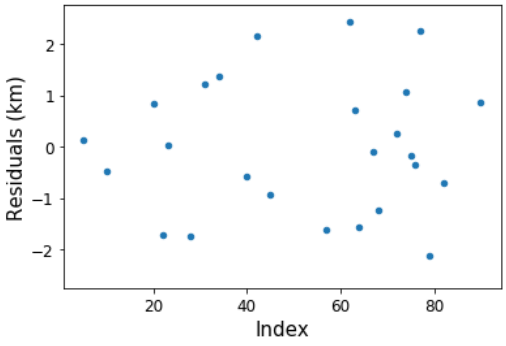
(iv) Weighted residuals vs. plane area



(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

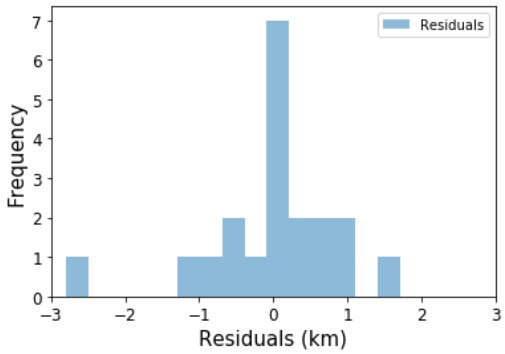
N/A

(vi) Weighted residuals vs. $\log_{\sqrt{2}}$ (UHs per hectare)

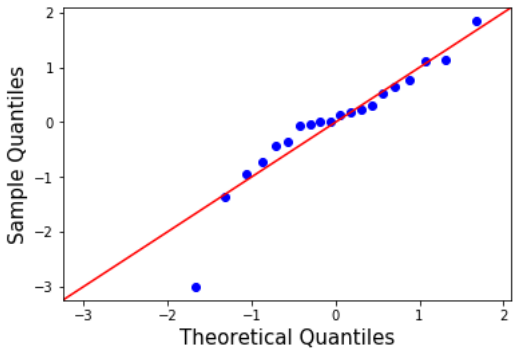


(vii) Residuals vs. order of observation

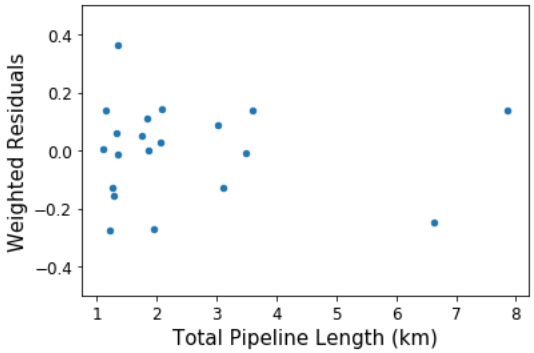
Figure E-17: OLS assumption check plots ('Non-Residential', 40 – 120 ha).



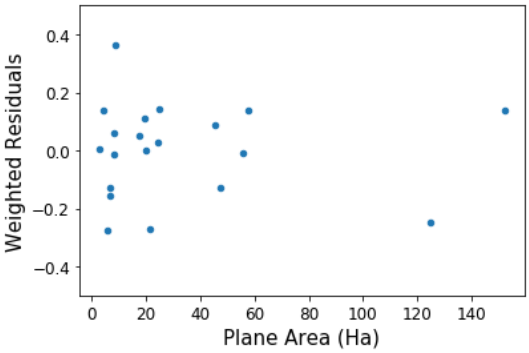
(i) Residuals histogram



(ii) Normal probability plot



(iii) Weighted residuals vs. total pipeline length



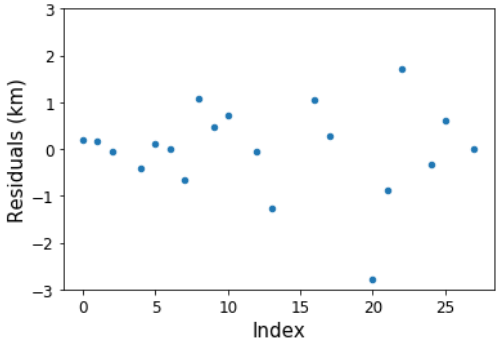
(iv) Weighted residuals vs. plane area

N/A

N/A

(v) Weighted residuals vs. $\sqrt{\text{mean relief}}$

(vi) Weighted residuals vs. $\log_{\sqrt{2}}$(UHs per hectare)



(vii) Residuals vs. order of observation

Figure E-18: OLS assumption check plots ('Large', 0 – 160 ha).

Appendix F
STUDY OUTCOME II: SETTING DIAMETER DISTRIBUTION CATEGORIES

Figure F-1 to Figure F-6 show scatter plots of the maximum nominal diameter versus plane area, and the total pipeline volume over length versus plane area for the 'General Residential', 'Low Income Residential', and 'Non-Residential and Large' land use categories. The plots were used to aid in setting area size category boundaries for the pipeline diameter distributions.

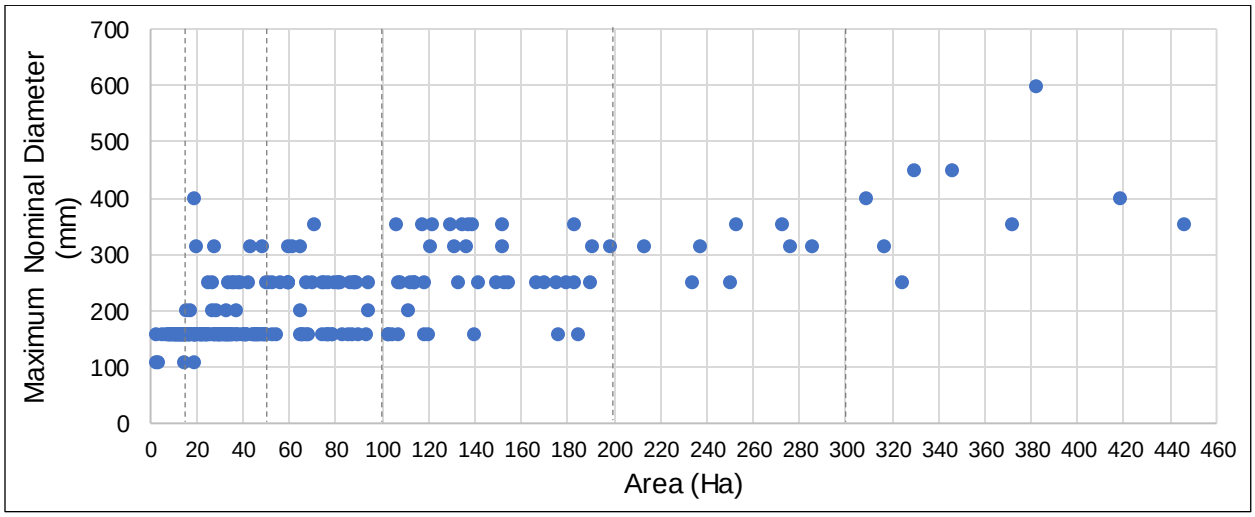


Figure F-1: Maximum nominal diameter vs. plane area ('General Residential').

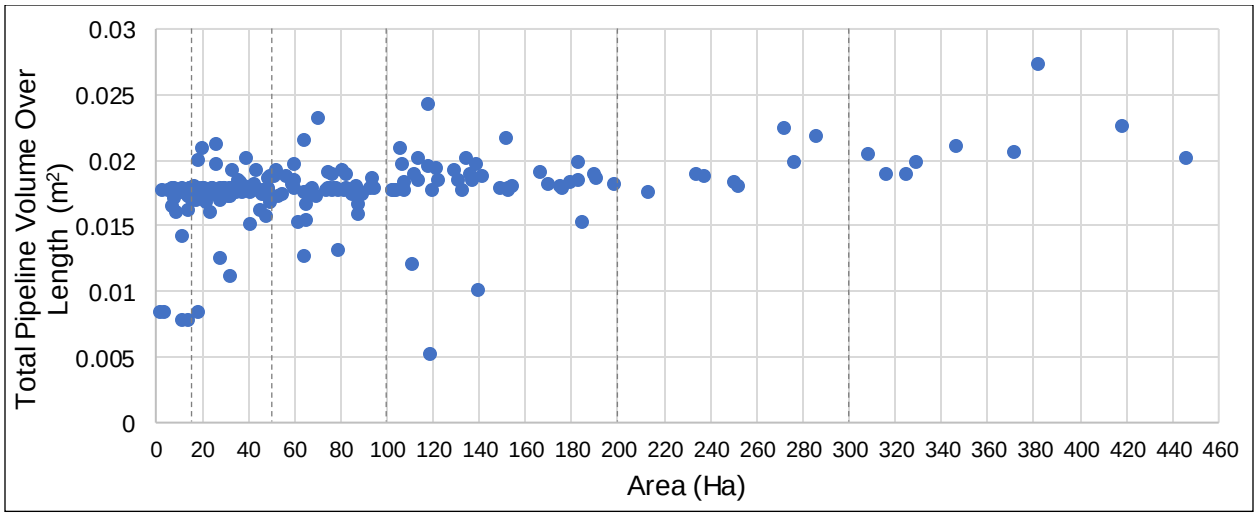


Figure F-2: Total pipeline volume over length vs. plane area ('General Residential').

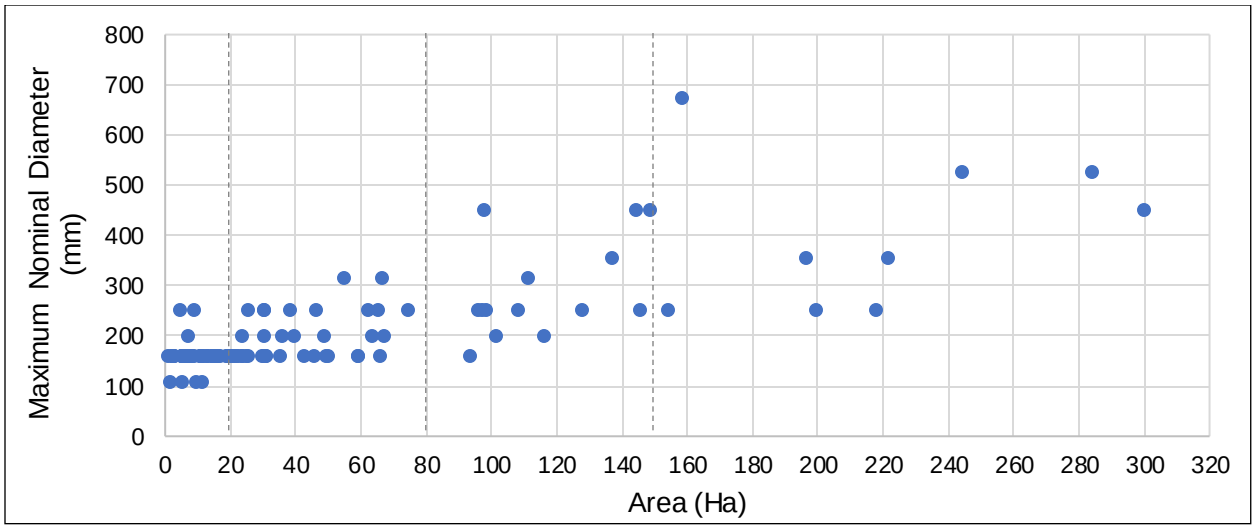


Figure F-3: Maximum nominal diameter vs. plane area ('Low Income Residential').

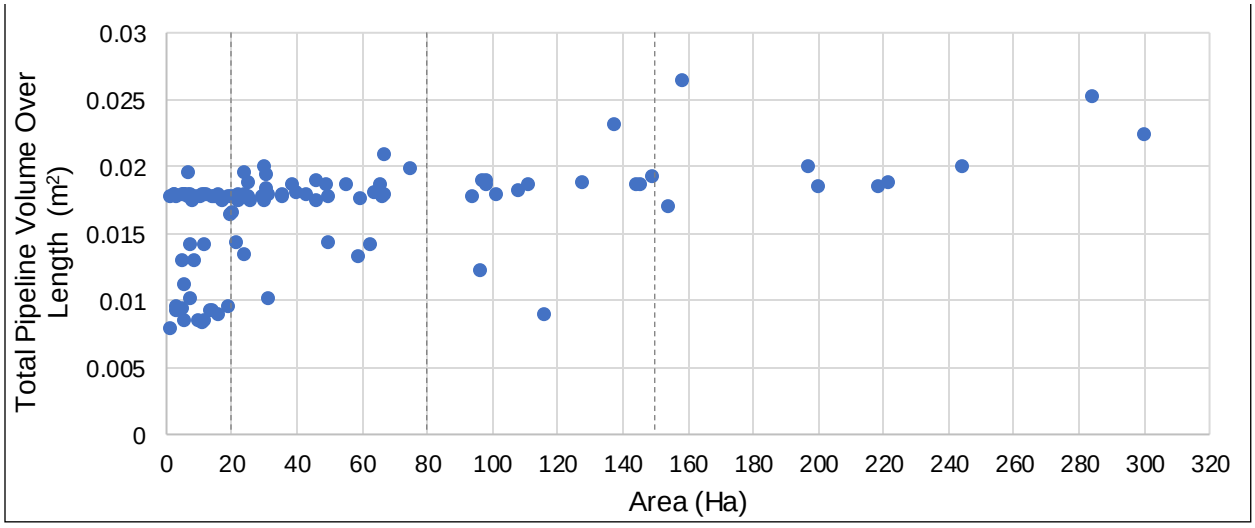


Figure F-4: Total pipeline volume over length vs. plane area ('Low Income Residential').

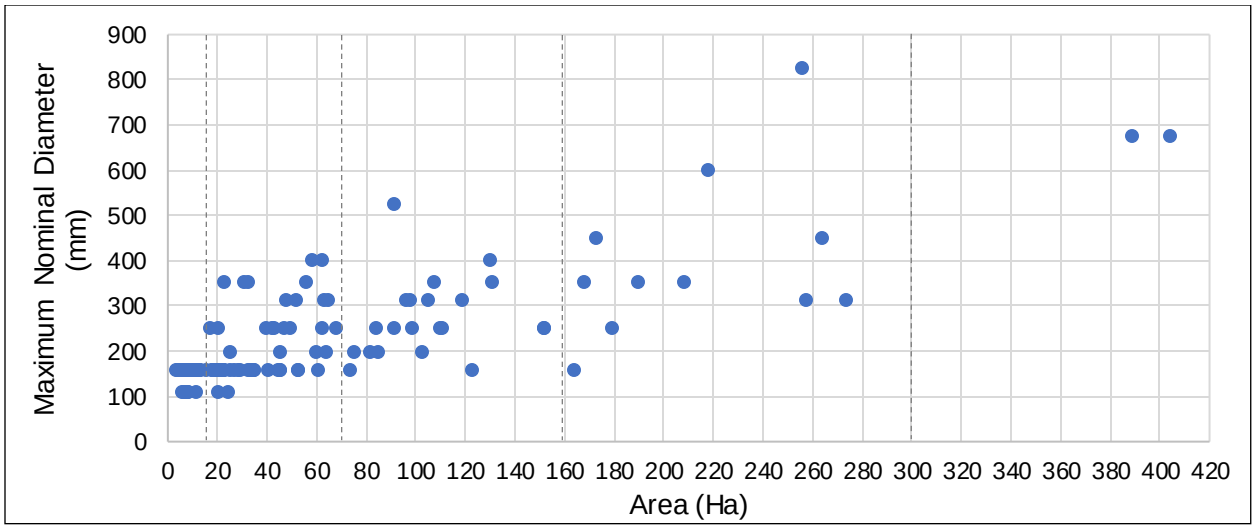


Figure F-5: Maximum nominal diameter vs. plane area ('Non-Residential and Large').

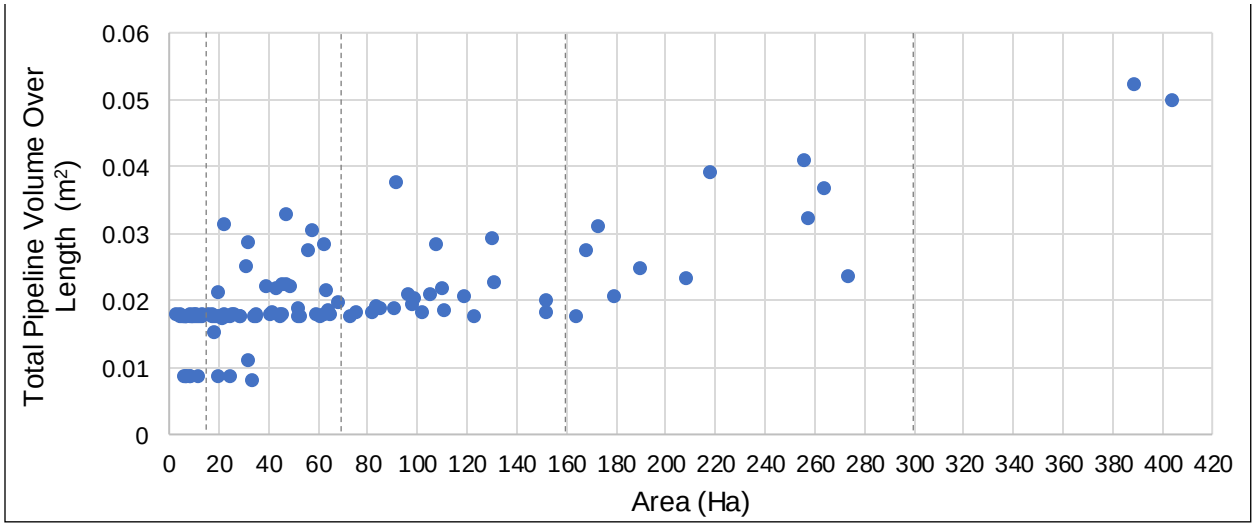


Figure F-6: Total pipeline volume over length vs. plane area ('Non-Residential and Large').

Appendix G

STUDY OUTCOME III: MANHOLE DISTRIBUTION DEVELOPMENT RESULTS

Table G-1 shows the regression model results from the investigation into which variables influence the average manhole frequency. The results were conflicting. Overall, variable combination A was favoured by the adjusted R^2 and BIC; and variable combination C appeared to be favoured by the log-likelihood and AIC. These results therefore did not provide convincing statistical evidence that the inclusion of any topography factor (as in variable combinations B, C and D) would measurably improve the estimates that could be made using only the total pipeline length (variable combination A).

Table G-1: Model results from Step 2: Significant quantitative variables.

Variable Combination		**A**	**B**	**C**	**D**
Independent Variables	Total pipeline length	x	x	x	x
	Total relief		x		
	Mean relief			x	
	Elevation standard deviation				x
General Residential	R^2 adjusted	**0.98**	0.96	0.96	0.96
	Log-likelihood	-971	-969	**-968**	-969
	AIC	1944	1944	**1942**	1944
	BIC	**1947**	1954	1951	1954
Low Income Residential	R^2 adjusted	**0.98**	0.96	0.96	0.96
	Log-likelihood	-457	-453	**-452**	**-452**
	AIC	916	912	**910**	911
	BIC	919	920	**918**	**918**
Non-Residential	R^2 adjusted	**0.98**	0.94	0.94	0.94
	Log-likelihood	-259	**-254**	**-254**	-256
	AIC	519	514	**513**	518
	BIC	522	521	**520**	524
Large	R^2 adjusted	**0.95**	0.90	0.90	0.91
	Log-likelihood	-89	**-88**	-89	**-88**
	AIC	**180**	182	183	181
	BIC	**181**	185	186	184

Appendix H
STUDY OUTCOME I: MODEL FORMULAE

Case A: All Variables Available

Model equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{x_2} + \beta_3 \log_{\sqrt{2}}(x_3)$

Table H-1: Variables for variable Case A.

Symbol	Variable	Unit	Calculation
y	Total pipeline length	km	-
x_1	Plane area	ha	-
x_2	Mean relief	m	Table 4-3 in Section 4.5.
x_3	UHs per hectare	Number/ha	Table D-1 in Appendix D.2

Table H-2: Regression coefficients for variable Case A.

Land Use Category	Area Size (ha)	β_0	β_1	β_2	β_3
General Residential	0 – 20	*-2.694*	*0.134*	*0.040*	*0.167*
		-1.845	**0.157**	**0.154**	**0.254**
		-0.996	0.180	*0.268*	*0.340*
	20 – 40	*-5.809*	*0.109*	*0.258*	*0.334*
		-4.189	**0.155**	**0.455**	**0.469**
		-2.569	*0.202*	*0.653*	*0.604*
	40 – 100	*-1.791*	*0.075*	*0.189*	*0.000*
		0.329	**0.102**	**0.530**	**0.000**
		2.448	*0.128*	*0.872*	*0.000*
	100 – 450	*-10.301*	*0.099*	*0.950*	*0.000*
		-6.214	**0.114**	**1.765**	**0.000**
		-2.128	*0.130*	*2.580*	*0.000*
Low Income Residential	0 – 40	*-4.180*	*0.169*	*0.112*	*0.172*
		-2.974	**0.187**	**0.244**	**0.297**
		-1.769	*0.205*	*0.376*	*0.422*
	40 – 300	*-27.043*	*0.134*	*0.144*	*0.949*
		-17.693	**0.153**	**0.884**	**1.962**
		-8.343	*0.171*	*1.624*	*2.974*
Non-Residential	0 – 40	*-0.845*	*0.064*	*0.009*	*0.069*
		-0.454	**0.083**	**0.142**	**0.114**
		-0.062	*0.102*	*0.274*	*0.160*
	40 – 120	*-2.974*	*0.034*	*0.522*	*0.000*
		-0.972	**0.060**	**0.885**	**0.000**
		1.029	*0.087*	*1.248*	*0.000*
Large	0 – 160	*0.635*	*0.029*	*0.000*	*0.000*
		0.961	**0.045**	**0.000**	**0.000**
		1.287	*0.062*	*0.000*	*0.000*
		Key:	*Lower confidence limit*	**Average**	*Upper confidence limit*

Case B: Area and Mean Relief Available

Model equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{x_2}$

Table H-3: Variables for variable Case B.

Symbol	Variable	Unit	Calculation
y	Total pipeline length	km	-
x_1	Plane area	ha	-
x_2	Mean relief	m	Table 4-3 in Section 4.5.

Table H-4: Regression coefficients for variable Case B.

Land Use Category	Area Size (ha)	β_0	β_1	β_2
General Residential	0 – 20	*-0.469*	*0.138*	*-0.060*
		-0.036	**0.161**	**0.069**
		0.397	*0.184*	*0.197*
	20 – 40	*-3.706*	*0.117*	*0.077*
		-1.659	**0.181**	**0.342**
		0.389	*0.244*	*0.607*
	40 – 100	*-1.791*	*0.075*	*0.189*
		0.329	**0.102**	**0.530**
		2.448	*0.128*	*0.872*
	100 – 450	*-10.301*	*0.099*	*0.950*
		-6.214	**0.114**	**1.765**
		-2.128	*0.130*	*2.580*
Low Income Residential	0 – 40	*-0.801*	*0.161*	*0.065*
		-0.299	**0.183**	**0.223**
		0.204	*0.205*	*0.380*
	40 – 300	*-3.708*	*0.121*	*-0.031*
		-0.347	**0.143**	**0.889**
		3.015	*0.165*	*1.809*
Non-Residential	0 – 40	*-0.683*	*0.061*	*-0.001*
		-0.176	**0.086**	**0.195**
		0.332	*0.110*	*0.390*
	40 – 120	*-2.974*	*0.034*	*0.522*
		-0.972	**0.060**	**0.885**
		1.029	*0.087*	*1.248*
Large	0 – 160	*0.635*	*0.029*	*0.000*
		0.961	**0.045**	**0.000**
		1.287	*0.062*	*0.000*
		Key:	*Lower confidence limit* / **Average** / *Upper confidence limit*	

Case C: Only Area Available

Model equation: $y = \beta_0 + \beta_1 x_1$

Table H-5: Variables for variable Case C.

Symbol	Variable	Unit	Calculation
y	Total pipeline length	km	-
x_1	Plane area	ha	-

Table H-6: Regression coefficients for variable Case C.

Land Use Category	Area Size (ha)	β_0	β_1
General Residential	0 – 20	*-0.043*	*0.139*
		0.166	**0.162**
		0.376	*0.185*
	20 – 40	*-2.645*	*0.121*
		-0.683	**0.192**
		1.279	*0.262*
	40 – 100	*0.977*	*0.074*
		2.725	**0.103**
		4.473	*0.132*
	100 – 450	*-2.066*	*0.106*
		0.841	**0.124**
		3.748	*0.141*
Low Income Residential	0 – 40	*0.077*	*0.168*
		0.328	**0.191**
		0.579	*0.213*
	40 – 300	*0.148*	*0.131*
		2.258	**0.152**
		4.367	*0.173*
Non-Residential	0 – 40	*-0.125*	*0.067*
		0.260	**0.094**
		0.645	*0.120*
	40 – 120	*-1.222*	*0.043*
		1.313	**0.079**
		3.848	*0.115*
Large	0 – 160	*0.635*	*0.029*
		0.961	**0.045**
		1.287	*0.062*
	Key:	*Lower confidence limit*	
		Average	
		Upper confidence limit	

Appendix I

STUDY OUTCOME I: MODEL PERFORMANCE RESULTS

This appendix contains the performance results of the total pipeline length models from Study Outcome I, for all three variable availability cases (A, B and C). Table I-1 provides the R^2, Table I-2 provides the MAPE and Table I-3 provides the 90% MAPE. It is recommended that, before implementing any model, the results in Table I-1, Table I-2 and Table I-3 should be considered by the user to establish whether the R^2 and MAPE are acceptable given the intended application. For example, the R^2 is low enough to indicate unreliable performance in the following models:

- 'General Residential' 20 – 40 ha, Case B and C
- 'General Residential' 40 – 100 ha, Case C
- 'Non-Residential' 0 – 40 ha, Case B and C
- 'Non-Residential' 40 – 120 ha, Case C

Table I-1: Model R^2 for variable cases A, B and C.

Land Use Category	Area Size (ha)	A		B		C	
		Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
General Residential	0 – 20	0.84	0.86	0.84	0.68	0.84	0.62
	20 – 40	0.80	0.91	0.51	0.65	0.45	0.46
	40 – 100	0.61	0.80	0.61	0.80	0.51	0.68
	100 – 450	0.87	0.94	0.87	0.94	0.81	0.89
Low Income Residential	0 – 40	0.91	0.93	0.87	0.90	0.85	0.89
	40 – 300	0.94	0.98	0.90	0.93	0.89	0.91
Non-Residential	0 – 40	0.81	0.60	0.68	0.17	0.60	0.11
	40 – 120	0.75	0.62	0.75	0.62	0.47	0.57
Large	0 – 160	0.64	0.04	0.64	0.04	0.64	0.04

Table I-2: Model MAPE (%) for variable cases A, B and C.

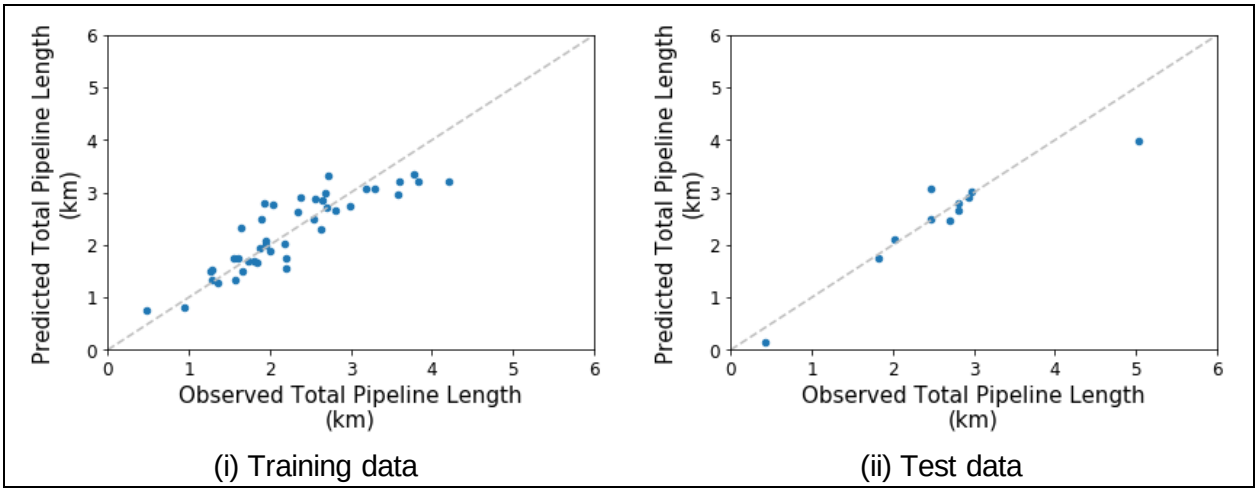
Land Use Category	Area Size (ha)	A		B		C	
		Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
General Residential	0 – 20	14.8	12.6	18.4	16.9	18.9	17.2
	20 – 40	12.6	9.3	17.2	19.2	20.0	19.6
	40 – 100	13.9	13.1	13.9	13.1	14.8	16.0
	100 – 450	13.4	9.6	13.4	9.6	15.3	14.2
Low Income Residential	0 – 40	19.9	17.3	24.8	16.1	25.6	32.7
	40 – 300	10.2	7.7	12.7	18.6	12.8	16.9
Non-Residential	0 – 40	25.2	22.0	31.1	44.6	31.2	49.2
	40 – 120	18.9	20.8	18.9	20.8	28.7	24.5
Large	0 – 160	35.0	30.6	35.0	30.6	35.0	30.6

Table I-3: Model 90% MAPE (%) for variable cases A, B and C.

Land Use Category	Area Size (ha)	A		B		C	
		Training Data	Test Data	Training Data	Test Data	Training Data	Test Data
General Residential	0 – 20	11.0	5.3	12.5	10.8	13.4	10.3
	20 – 40	10.6	6.5	13.4	10.9	16.6	12.2
	40 – 100	11.1	11.9	11.1	11.9	12.1	12.7
	100 – 450	10.9	7.7	10.9	7.7	11.6	9.2
Low Income Residential	0 – 40	14.7	11.2	14.6	13.9	13.7	13.5
	40 – 300	8.3	7.0	9.9	16.4	9.7	14.4
Non-Residential	0 – 40	19.1	18.6	20.6	32.8	20.2	33.8
	40 – 120	15.1	15.8	15.1	15.8	20.0	20.5
Large	0 – 160	22.0	23.9	22.0	23.9	22.0	23.9

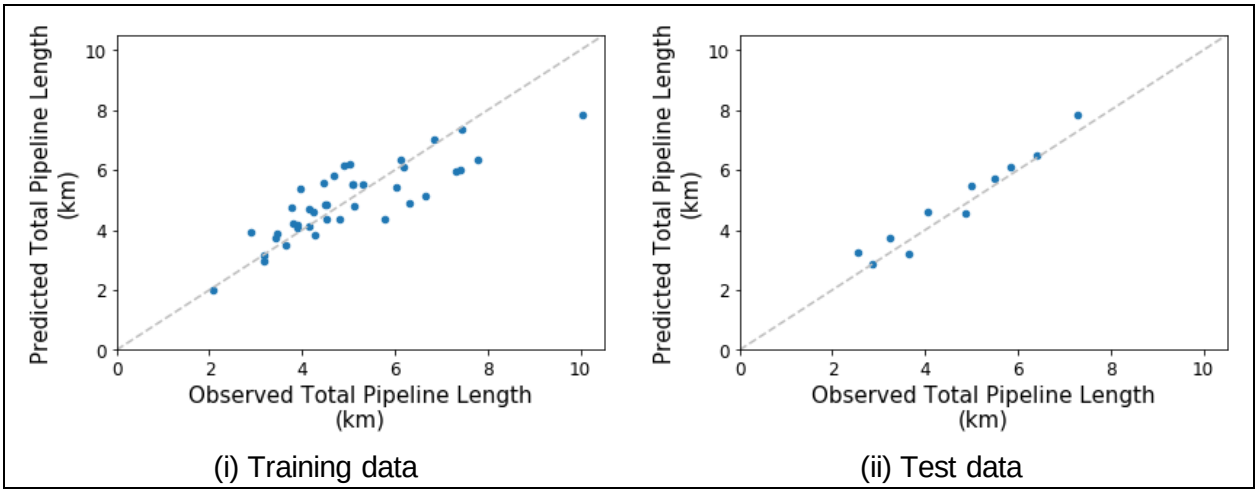
Appendix J
STUDY OUTCOME I: CASE A MODEL PERFORMANCE PLOTS

Figure J-1 to Figure J-9 illustrate the strength of the nine Case A models in the form of scatter plots of the predicted versus observed total pipeline length. The strength of the scatter correlates to the R^2. The range of values on the axes differs between models, therefore while the scatter strength of two models may be visually similar, the size of the residuals might differ.



(i) Training data (ii) Test data

Figure J-1: Predicted vs. observed total pipeline length ('General Residential', 0 – 20 ha).



(i) Training data (ii) Test data

Figure J-2: Predicted vs. observed total pipeline length ('General Residential', 20 – 40 ha).

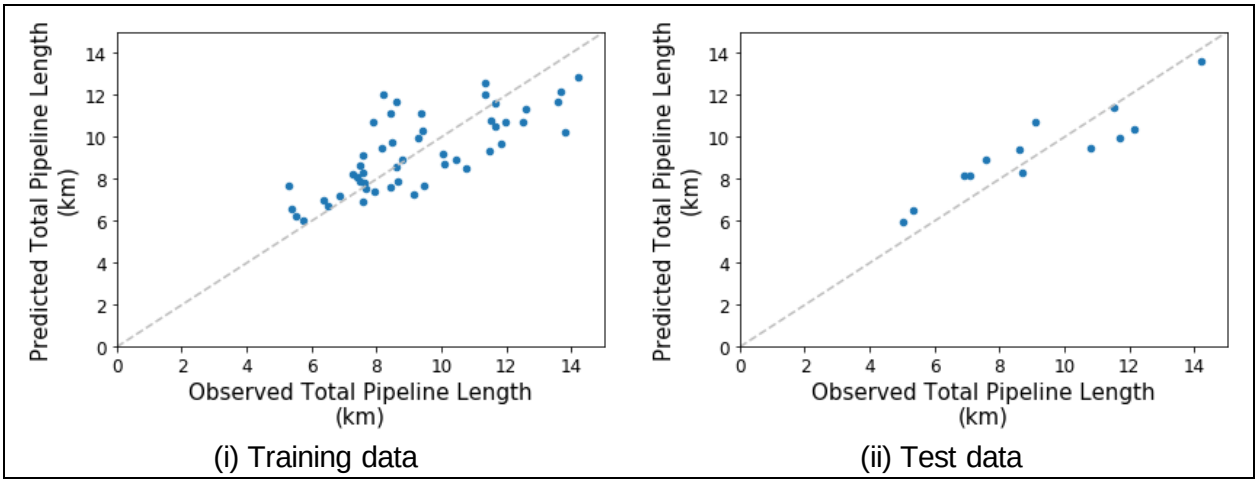


Figure J-3: Predicted vs. observed total pipeline length ('General Residential', 40 – 100 ha).

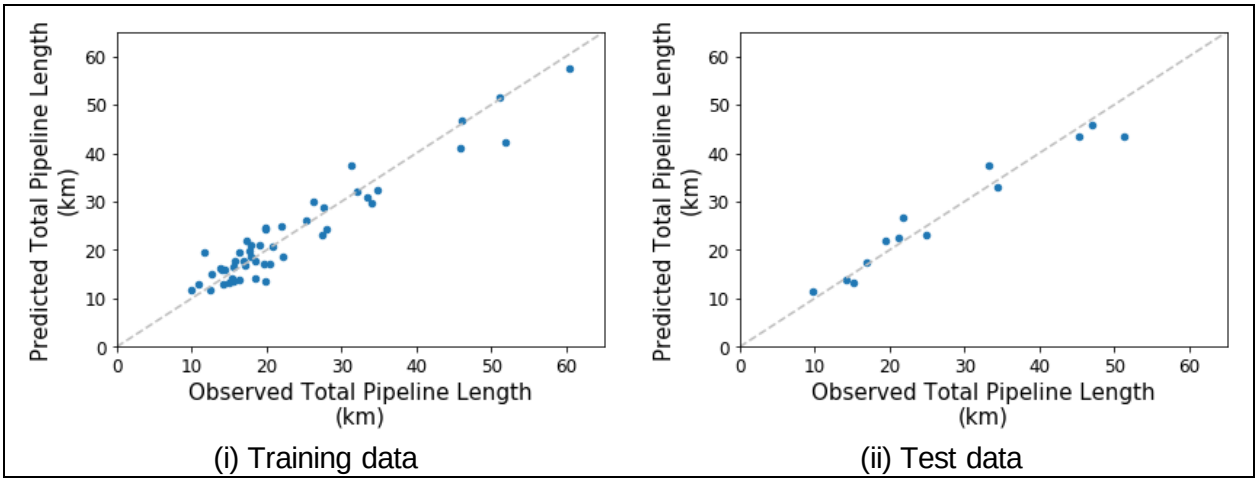


Figure J-4: Predicted vs. observed total pipeline length ('General Residential', 100 – 450 ha).

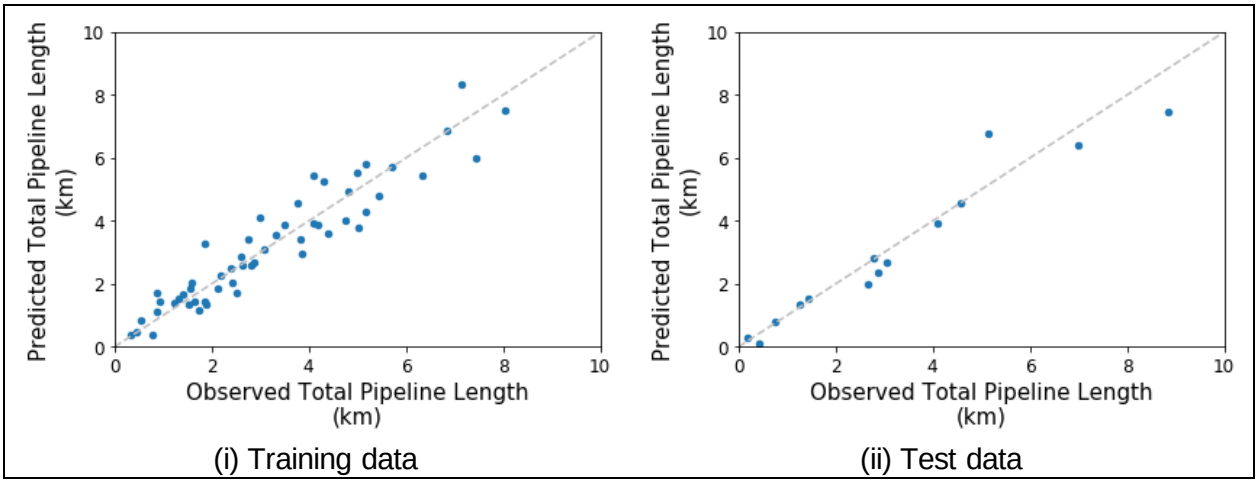


Figure J-5: Predicted vs. observed total pipeline length ('Low Income Residential', 0 – 40 ha).

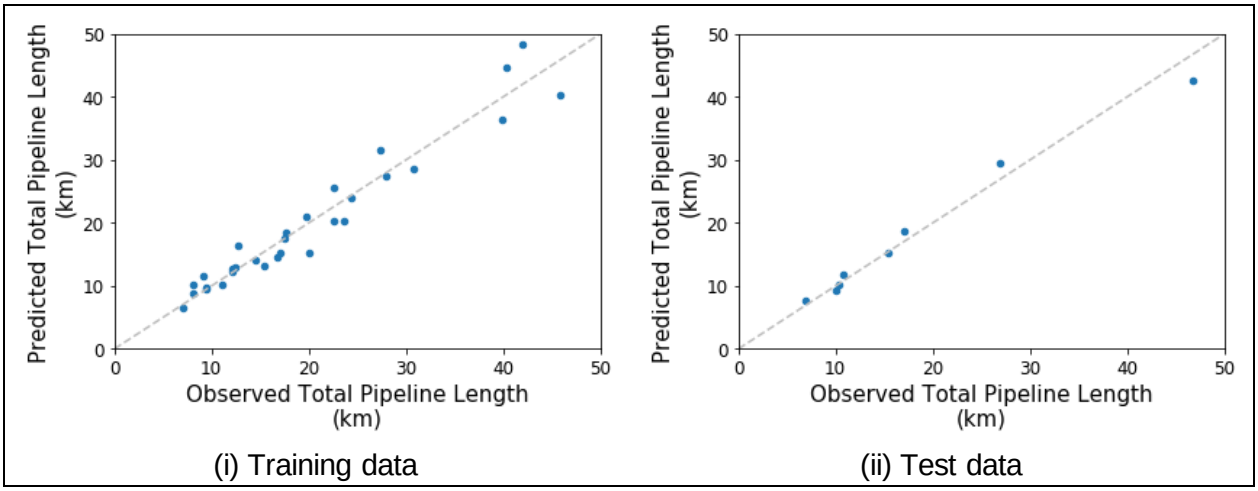


Figure J-6: Predicted vs. observed total pipeline length ('Low Income Residential', 40 – 300 ha).

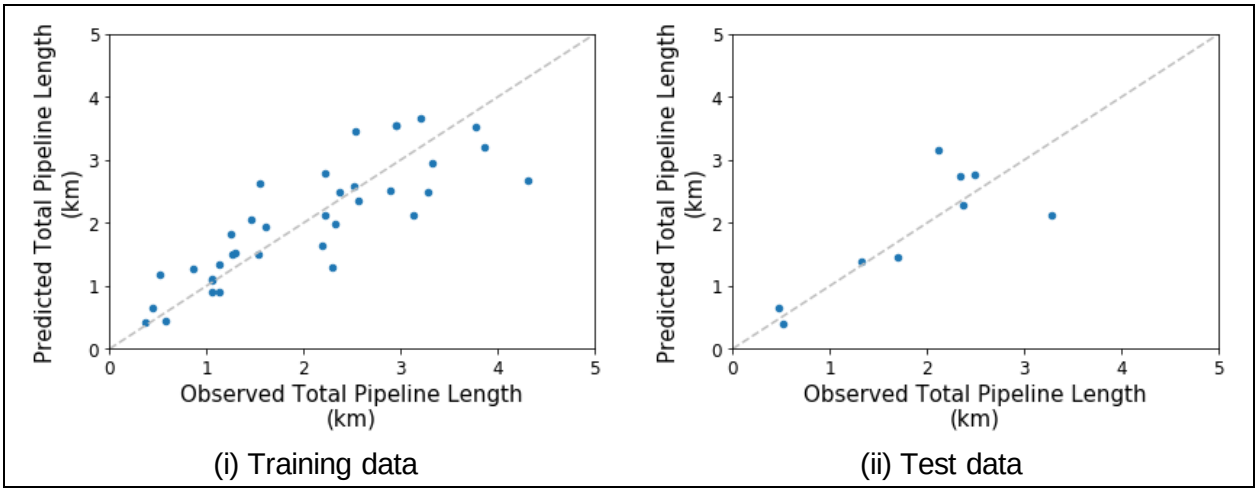


Figure J-7: Predicted vs. observed total pipeline length ('Non-Residential', 0 – 40 ha).

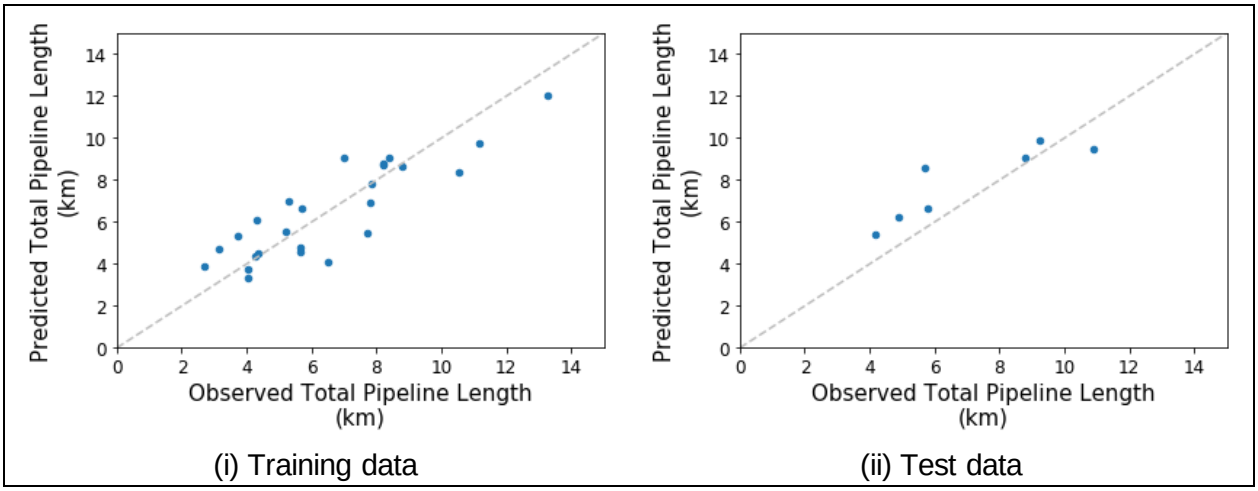


Figure J-8: Predicted vs. observed total pipeline length ('Non-Residential', 40 – 120 ha).

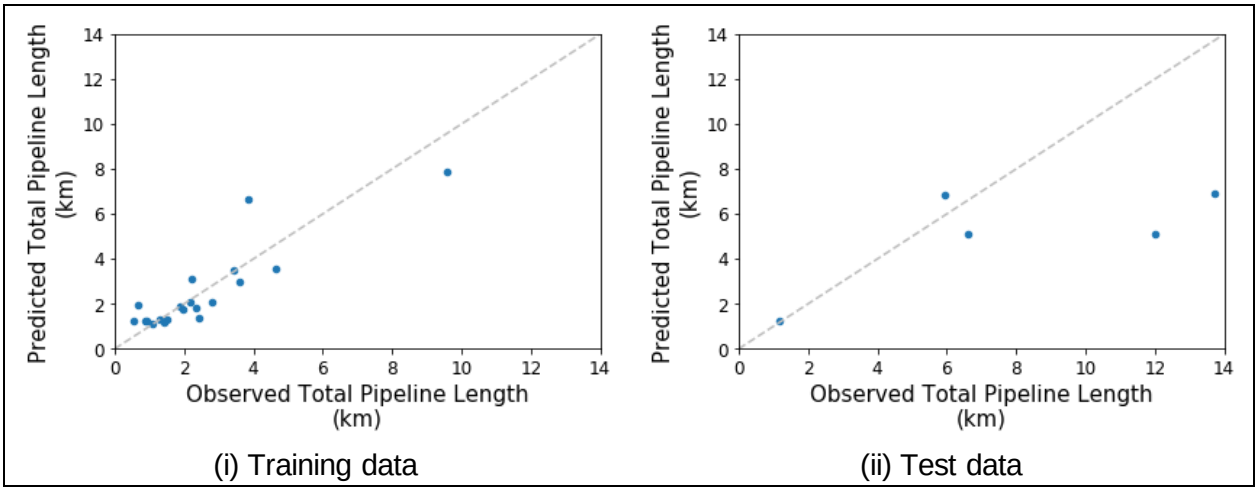


Figure J-9: Predicted vs. observed total pipeline length ('Large', 0 – 160 ha).

Appendix K
APPLICATION EXAMPLE

This appendix contains an example of how the infrastructure estimation tool developed in this study can be applied to a service zone. This example aims to illustrate how to use the tool, but also how the results can be interpreted flexibly based on the application requirements.

Consider a proposed development of 22.07 hectares. The planned number of connected users and the associated wastewater flow production (excluding infiltration) are presented in Table K-1. The unit hydrographs (UHs) were assigned according to Table D-1 in Appendix D.2. Based on a digital elevation model (DEM) of the service zone, the mean elevation is 473.6 MASL, and the elevation of the proposed network endpoint (or lowest convergence point) will be at approximately 455.5 MASL. The network endpoint is approximately 50 m away from the bulk line to which the network will be connected. Furthermore, municipal specifications stipulate a minimum nominal diameter of 160 mm. The service zone is positioned in a region that is not underlain by dolomite, therefore according to the DHS (2019) guidelines, it can be assumed that pipes less than 400 mm in diameter will be uPVC, and greater than 400 mm in diameter will be concrete. The developer requires an estimate of the expected sewer pipeline infrastructure for early-stage cost estimation.

Table K-1: Connected users and flow production per land use.

Land Use	UH Unit	Number of UHs	PDDWF (kL/d)
Low density residential	erf	75.0	52.5
Medium density residential	erf	70.0	42.0
Flats	unit	50.0	20.1
Business/ commercial	100 m^2 floor	9.5	5.0
Total		**204.5**	**119.6**

First, the dominant land use category from Table 4-4 (Chapter 4) that best describes the service zone must be determined. As noted in Table 4-3, the dominant land use category for this study was determined as the one with the greatest contribution to the total PDDWF. However, it was also noted that in cases where flow information is not available, considering the contribution to the total UH count would probably yield the same answer in the majority of cases. For illustration purposes, both methods are presented in Table K-2. Both methods indicate that the dominant land use category is ‘General Residential’.

Table K-2: Land use category representation according to PDDWF and UH contribution.

Land Use Category	PDDWF Contribution (%)		UH Contribution (%)	
General Residential	$\frac{52.5+42+20.1}{119.6}\times 100 =$	95.8	$\frac{75+70+50}{204.5}\times 100 =$	95.4
Low Income Residential	$\frac{0}{119.6}\times 100 =$	0	$\frac{0}{204.5}\times 100 =$	0
Non-Residential	$\frac{5.0}{119.6}\times 100 =$	4.2	$\frac{9.5}{204.5}\times 100 =$	4.6
Large	$\frac{0}{119.6}\times 100 =$	0	$\frac{0}{204.5}\times 100 =$	0
Total		**100**		**100**

The next step is to quantify the three required input variables for the infrastructure estimation tool. This step is presented in Table K-3.

Table K-3: Input variables for infrastructure estimation tool.

Input Variable	Calculation	Value	Unit
Plane area	-	22.07	ha
Mean relief	473.6 - 455.5	18.10	m
UHs per hectare	204.5 ÷ 22.07	9.27	UH/ha

The infrastructure components, namely the pipeline length per diameter and the number of manholes, can now be estimated. For the total pipeline length, the relevant model category is 'General Residential', 20 – 40 ha. Using the Case A model defined in Equation 9-1 (Chapter 9), with the 'Average' regression coefficients from Table 9-3, yields Equation K-1. The values from Table K-3 are substituted and the equation is solved to obtain the total pipeline length.

$$Total\ Pipeline\ Length = -4.189 + 0.155(Plane\ area) + 0.455\sqrt{Mean\ relief} + 0.469\log_{\sqrt{2}}(UHs\ per\ hectare) \quad \text{K-1}$$

$$Total\ Pipeline\ Length = -4.189 + 0.155(22.07) + 0.455\sqrt{18.10} + 0.469\log_{\sqrt{2}}(9.27)$$

$$Total\ Pipeline\ Length = 4.180\ km$$

It is noted that this total pipeline length of 4.180 km represents the average expected value. That is, there is a 50% chance that the true value will be greater than this, and a 50% chance that the true value will be less than this. Furthermore, from Table 9-5 (Chapter 9), the mean absolute percentage error (MAPE) for the particular model used is in the order of 10%. The MAPE indicates that the true value can *on average* be expected to deviate from the estimate by roughly 10%. This could be accounted for if a more conservative estimate were required – for example, by increasing the estimate by 10%. Nonetheless, for this example, a simple estimate is required.

To disaggregate the total pipeline length into lengths per diameter, the 'General Residential', 15 – 50 ha diameter distribution category is applicable. Table K-4 presents the distribution, the diameter disaggregation, and the adjustment thereof for the minimum-diameter specification.

Table K-4: Disaggregation of total pipeline length to length per diameter.

Nominal Diameter (mm)	Proportion of Total Pipeline Length (%)	Pipeline Length (m)	Adjusted Pipeline Length (m)
110	4.0	167	0
160	94.6	3955	4122
200	0.4	17	17
250	0.8	33	33
315	0.2	8	8
Total	**100**	**4180**	**4140**

In addition to the network pipes in Table K-4, the 50 m pipeline section connecting the network endpoint to the bulk pipeline must be accounted for. Depending on the information available, the diameter of this section can be estimated in different ways. Ideally, if the total expected PDDWF for the service zone is available as it is in this example, then the design flow can be estimated by first adding the infiltration associated with the pipes in Table K-4 to the PDDWF, then converting this to the IPWWF flow by applying a suitable peak hour factor and accounting for the required spare capacity (DHS, 2019). Using the estimated IPWWF, the associated pipe diameter required to convey this flow can be determined. If the total expected PDDWF is not known, then a logical estimation would have to be made. For the purposes of this example, presume that by assuming a suitable infiltration rate, peak hour factor, and spare capacity, the pipe diameter required to convey the associated IPWWF is 355 mm. Then, since all of the pipes in the network are less than 400 mm in diameter, it is assumed that only uPVC pipes will be used.

Finally, the number of manholes can be determined based on the total pipeline length (including the additional 50 m). The 'Residential', 20 – 50 ha distribution category is applicable, which stipulates an average of 21.3 manholes/km. Therefore the total number of manholes is calculated as follows:

$$Number\ of\ Manholes = 21.3\ \times Total\ Pipeline\ Length$$

$$Number\ of\ Manholes = 21.3\ \times (4.180 + 0.05)$$

$$Number\ of\ Manholes = 90$$

In summary, the proposed development can be expected to require sewer pipeline infrastructure of the order of:

- 4122 m of 160 mm uPVC pipe
- 17 m of 200 mm uPVC pipe
- 33 m of 250 mm uPVC pipe
- 8 m of 315 mm uPVC pipe
- 50 m of 355 mm uPVC pipe
- 90 manholes.

It is noted that the actual sewer infrastructure will certainly deviate from the above, but that the above will enable a cost estimation that is tailored to the proposed development.